Conformal Refusal For Quantized Small Language Models: Conditional Risk Control Under Resource Constraints

Mrs. Ganga Naidu[1], Julian Fidal Castro S[2], Aarya A[3], Karthick Subramanian K[4], Logeshwaran A[5]
[1]*Assistant Professor, K. Ramakrishnan College of Technology, Trichy, Tamil Nadu, India*
[2,3,4,5]*Student, Department of Artificial Intelligence and Data Science, K. Ramakrishnan College of Technology, Trichy, Tamil Nadu, India*

Abstract*—Small language models present a feasible choice for employing artificial intelligence in certain devices that have limited resources. However, the estimates made by them are not necessarily reliable. This research work studies a conformal refusal framework for a quantized small language model that aims to keep the number of errors in predictions not refused. The framework has been tested using Qwen2.5-1.5B-Instruct with FP16 and 4-bit NF4 quantization on the configuration of TruthfulQA having 817 questions with multiple choices. An initial likelihood-based score indicating confidence turned out to be strongly dependent on the number of answer choices, so a letter-based in-context comparative scoring method was used instead. We consider three approaches to calibration: Conformal Risk Control, Learn-then-Test (with the Bonferroni correction), and Fixed Sequence Testing. We evaluate these methods against fixed-threshold and prompted-refusal baselines, tested on 100 resamples. The results show that the Bonferroni approach offers empirical coverage rates of approximately 98.9 percent for FP16 and 98.3 percent for NF4, exceeding the expected nominal coverage of 90 percent with respect to an error target of 0.50. Reliable certification was obtained only at error targets of 0.50 and above. For the target of 0.50, the feasibility of threshold dropped from 93 percent for FP16 to 58 percent for NF4. Overall, the results highlight how conformal calibration can be useful for selective refusal of predictions in the small quantized language model.

Index Terms*—Small language models, conformal prediction, selective prediction, risk control, quantization, abstention, uncertainty calibration, edge artificial intelligence.

I. INTRODUCTION

Small language models (SLMs) have become an important option for deploying artificial intelligence in environments with limited computational resources. Their relatively compact size makes them suitable for local inference on consumer devices and edge hardware, reducing dependence on cloud infrastructure. Research on small language models highlights their potential for resource-efficient applications [1]. Quantization further reduces the memory required to store model parameters, enabling language models to operate within restricted hardware budgets [2]. However, reducing model size and memory consumption does not necessarily ensure reliable predictions. Confidence estimation and uncertainty quantification remain important challenges for language models, as high confidence does not always correspond to correct answers [3].
Selective prediction addresses this problem by allowing a model to abstain when its confidence is insufficient [4]. Instead of answering every question, a model can provide an answer only when it satisfies a specified confidence criterion. However, conventional confidence scores may not reliably distinguish correct predictions from incorrect ones. Research on neural-network calibration emphasizes the importance of evaluating the relationship between predicted confidence and actual correctness [5]. This challenge is particularly relevant to multiple-choice question answering, where a model must select the correct answer from several alternatives. TruthfulQA provides a benchmark designed to evaluate whether language models reproduce common misconceptions and falsehoods [6].

Conformal methods offer a statistical approach to uncertainty management and prediction-risk control. Conformal Risk Control (CRC) provides a framework

for controlling expected risk under suitable assumptions [7], while Learn-then-Test (LTT) formulates risk control as a multiple-hypothesis testing problem [8]. Distribution-free risk-controlling prediction methods provide further foundations for finite-sample risk guarantees [9]. Nevertheless, controlling the joint probability that a model answers incorrectly does not automatically guarantee a low error rate among the questions it actually answers. Conformal abstention has been explored for language models [10], motivating further investigation into statistical refusal mechanisms that directly target conditional risk.

This study investigates whether conditional-risk calibration can support reliable refusal decisions in a quantized small language model operating under resource constraints. We evaluate Qwen2.5-1.5B-Instruct in FP16 and 4-bit NF4 configurations using the TruthfulQA MC1 dataset, which contains 817 multiple-choice questions. An initial likelihood-based confidence method was found to be strongly associated with the number of answer choices and almost uncorrelated with correctness. Therefore, a letter-based in-context comparative scoring method was adopted to obtain a more informative confidence signal.

Three calibration approaches are evaluated: CRC as a joint-risk baseline, Bonferroni-corrected LTT as the primary conditional-risk procedure, and Fixed Sequence Testing (FST) as an exploratory alternative. Their performance is compared across multiple conditional-risk targets and 100 calibration-test resamples, alongside fixed-threshold and prompted-refusal baselines.

The main contributions of this study are: (1) evaluating confidence estimation methods for multiple-choice prediction under a small language model; (2) distinguishing joint-risk control from conditional-risk certification; and (3) investigating how 4-bit NF4 quantization affects threshold feasibility and empirical certification performance compared with FP16.

The findings identify the risk levels at which certification was feasible in the evaluated setting and highlight limitations associated with calibration data availability and confidence-score reliability.

II. LITERATURE SURVEY

A. Small Language Models and Quantization

Small language models are increasingly studied as practical alternatives for applications with limited memory and computational resources [1]. Quantization is one approach to reducing model memory requirements by representing parameters with lower numerical precision. Dettmers et al. introduced QLoRA, which combines low-bit quantization with memory-efficient fine-tuning techniques for large language models [2]. Although this work provides relevant context for quantized language models, its primary focus is fine-tuning rather than the specific inference-time NF4 implementation evaluated in this study. The reliability of confidence estimates under quantization therefore remains an important consideration for resource-constrained deployment.

B. Uncertainty Quantification and Selective Prediction

Reliable uncertainty estimation is essential when a model must decide whether to provide an answer or abstain. Selective classification studies this decision by allowing a prediction system to reject uncertain inputs rather than always producing a prediction [4]. Guo et al. examined the calibration of modern neural networks and demonstrated that predictive confidence can differ from empirical accuracy [5]. Surveys of uncertainty quantification in language models further discuss the challenges of estimating uncertainty from model outputs [3].

These studies motivate evaluating confidence scores according to their relationship with correctness rather than assuming that a high score necessarily indicates a reliable answer.

C. TruthfulQA and Multiple-Choice Evaluation

TruthfulQA was introduced by Lin et al. to measure whether language models reproduce common human misconceptions and falsehoods [6]. Its multiple-choice MC1 configuration provides a suitable setting for evaluating answer selection and confidence-based refusal.

In the present study, the dataset contains 817 questions, each with one correct answer. The evaluation also examines whether the confidence score is influenced by the number of answer choices, since a score that primarily reflects question structure may be unsuitable for risk-based refusal.

D. Conformal Risk Control and Learn-then-Test

Conformal Risk Control provides a framework for controlling expected prediction risk under appropriate assumptions [7]. However, the guarantee associated with a particular loss function must match the quantity that the application intends to control. For selective question answering, joint risk measures the probability that the model both answers and makes an error, whereas conditional risk measures the probability of error given that an answer is provided. A bound on joint risk does not, by itself, guarantee the desired conditional-risk bound.

Learn-then-Test provides an alternative framework that evaluates candidate decision rules using statistical hypothesis tests while accounting for multiple comparisons [8]. Distribution-free risk-controlling prediction methods provide related theoretical foundations for finite-sample risk control [9]. In this study, Bonferroni-corrected LTT is used as the primary method because the target is conditional risk among answered questions. Fixed Sequence Testing is also evaluated as a less conservative alternative, but its suitability depends on the relevant ordering assumptions holding in practice.

E. Conformal Methods for Language Models and Research Gap

Prior work has investigated conformal abstention for large language models, exploring how statistical risk-control methods can support decisions to withhold potentially unreliable answers [10]. Other research has examined conformal prediction for multiple-choice question answering [11] and extended conformal techniques to language-model outputs more generally [12]. These studies establish a foundation for applying distribution-free uncertainty-control methods to language models.

However, the present study focuses specifically on conditional-risk certification for a small language model under low-bit quantization and limited edge-hardware resources. It compares FP16 and 4-bit NF4 configurations, evaluates alternative confidence-scoring methods, and examines the feasibility of statistical threshold certification across multiple risk targets. This focus distinguishes the study from work that considers conformal prediction more generally and provides an empirical investigation of how quantization affects conditional-risk calibration in the evaluated setting.

III. METHODOLOGY

A. Problem Formulation

Let x denote an input question and let y denote the model's predicted answer. The system produces either an answer or a refusal. A confidence score $s(x)$ is used to determine whether the prediction should be accepted.

For a threshold λ, the decision rule is

$$D_\lambda(x) = \begin{cases} \text{answer}, & s(x) \geq \lambda, \\ \text{refuse}, & s(x) < \lambda. \end{cases}$$

The primary objective is to control the probability that an accepted prediction is incorrect:

$$R_{\text{cond}}(\lambda) = P(\text{Incorrect} \mid \text{Not Refused}).$$

This conditional risk is different from the joint risk

$$R_{\text{joint}}(\lambda) = P(\text{Not Refused} \cap \text{Incorrect}).$$

The two quantities satisfy

$$R_{\text{joint}}(\lambda) = P(\text{Not Refused}) R_{\text{cond}}(\lambda).$$

Therefore, controlling joint risk alone does not directly guarantee the desired conditional error rate, particularly when the refusal rate is high.

B. Experimental System

Qwen2.5-1.5B-Instruct is tested in 2 inference modes—FP16 and 4-bit NF4 quantization. The experiments are conducted locally on the NVIDIA RTX 3050 Laptop GPU with 6 GB VRAM using Python, PyTorch, CUDA, and bitsandbytes without using cloud services.

In the FP16 configuration, the amount of memory used onboard by the model is about 3.09 GB. Meanwhile, in NF4, the equivalent number is equal to approximately 1.16 GB. The difference is roughly 2.7 times less. It is important to note that bitsandbytes NF4 double quantization is adopted in 4-bit implementation, which is different from GGUF-based INT4 alternatives.

C. Dataset and Data Preparation

The study employed the TruthfulQA MC1 dataset which is composed of 817 questions, with each question being equipped with either 2 to 13 choices, containing a single correct answer in each case. This dataset was chosen, as it aids in exploring the candidates' susceptibility to misconceptions.

The raw dataset provided the right answer for each of the questions at index zero. To avoid associating this index with the success of the prediction procedure, the answer options were shuffled in predetermined order based on the question ID number used as the random seed. The same shuffling of the alternatives was used in the case of FP16 and NF4 experiments.
In each case, the total of 817 questions were divided randomly into calibration and test parts making sure that both parts were of equal size. This splitting was repeated 100 times.

D. Confidence Estimation
Two confidence estimation approaches were investigated.

1) Likelihood-Based Confidence
In the beginning, the sequence log-likelihood of each possible answer was determined.
Nonetheless, this certainty measurement was significantly influenced by the number of options. The correlation with the number of options was nearly -0.93, while the correlation with accuracy was nearly 0.03. This truly suggests that the measurement primarily reflected competition among alternatives rather than effectively distinguishing right from wrong answers.

2) Letter-Based Comparative Confidence
Therefore, a different method was employed where all the options were displayed together in one prompt and labeled with letters like A, B, C, and D. The task of the algorithm was to choose the letter corresponding to the answer.
The logits for the letters belonging to the valid answer were transformed into probabilities and subsequently the probability of the letter chosen was used as the confidence level. This method resulted in a correlation of about 0.30 in the connection between confidence and correctness for FP16, while for NF4, the correlation was found to be at an approximate value of 0.32.
Both correlations are found to be statistically significant at $p<10{-}17$. However, there is still a residual correlation of around −0.23 with the number of choices.
Thus, the letter-based technique was then implemented in the following refusal and calibration tests.

E. Refusal Mechanism
At any given confidence, λ, the model either accepts its predicted answer or does not answer. Increasing λ means a more selective system that accepts only predictions with sufficiently high confidence.
The calibration challenge, thus, is to find a threshold such that a given conditional error target α can be met while accepting as many predictions as possible.

The primary calibration target is
$P(\text{Incorrect} \mid \text{Not Refused}) \leq \alpha$.

Seven target values were evaluated:
$\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60\}$.
The wider range was introduced because the initial experiments found no usable certified threshold at the tighter targets of 0.05, 0.10, and 0.20.

F. Calibration Procedures
Three calibration approaches were evaluated.
1) Conformal Risk Control
The first approach uses Conformal Risk Control (CRC) as a baseline. Its calibration objective controls the joint risk
$P(\text{Not Refused} \cap \text{Incorrect})$,
rather than directly controlling the desired conditional risk. The CRC threshold was therefore treated as a baseline rather than as the primary solution to the conditional-risk problem.

2) Bonferroni-Corrected Learn-then-Test
The main conditional-risk method involves using the Learn-then-Test procedure in conjunction with Bonferroni correction. There are 30 potential confidence thresholds in this analysis. The decision is made after each potential threshold has been evaluated using the conditional-risk criterion by means of a Hoeffding-Bentkus concentration bound. Given more than one confidence threshold used, the probability of making an error, $\delta=0.10$, is adjusted by applying a Bonferroni method. This system does not need to present a monotonous dependence between the confidence threshold and the conditional risk observed. Therefore, it can be considered the most reliable method.

3) Fixed Sequence Testing
As an alternative approach, Fixed Sequence Testing (FST) was tested out. The candidate thresholds were

evaluated based on their severity and the tests were immediately put to a halt after the first failure. The method circumvents the complete Bonferroni penalty if the risk in the population tends to be monotonous over the validated thresholds. However, bootstrapping has proven that the completely monotonic risk curve had only emerged for around 14–18% of the bootstraps that were generated. Hence, FST is regarded as an exploratory method and not a major certifying technique.

G. Reliability and Coverage Evaluation

Each calibration and test experiment uses a fresh approximately 50/50 calibration-test split. The complete evaluation consists of 100 independent resampling trials for each experimental configuration.

Trials are categorized according to the number of test examples accepted after calibration:

- No data: $n_{\text{found}} = 0$
- Inconclusive: $1 \leq n_{\text{found}} < 20$
- Reliable: $n_{\text{found}} \geq 20$

For trials with sufficient accepted examples, empirical coverage is calculated as

$$\text{Coverage} = \frac{\{\text{trials with achieved risk} \leq \alpha\}}{n_{\text{found}}}.$$

The nominal confidence level for the conditional-risk calibration procedures is 90 percent, corresponding to $\delta = 0.10$.

H. Baselines

Two non-conformal baselines are included for comparison:

1. Fixed confidence threshold: predictions are accepted when the confidence score is at least 0.5.
2. Prompted refusal: the model is explicitly instructed to refuse when it is uncertain.

These baselines allow the effect of statistical calibration to be compared against simple thresholding and model-generated refusal behaviour.

IV. EXPERIMENTAL RESULTS

A. Overall Results

The experiments studied the effectiveness of different configurations of Qwen2.5-1.5B-Instruct such as FP16 and 4-bit NF4 configurations. The assessment of all configurations was carried out based on 100 random calibration-test resamples.

Outcomes indicate a difference observed in controlling joint risk as compared to controlling conditional risk. While CRC helps achieve joint risk that is fairly close to the targeted value, problems arise with the conditional error being significantly high. On the other hand, the Learn-then-Test technique that uses Bonferroni correction allows one to use empirical control of conditional risk provided the calibration evidence is available.

B. Conformal Risk Control Baseline

Table I summarizes the CRC results.

TABLE I CRC RESULTS FOR FP16 AND NF4

α	FP16 Refusal	FP16 Joint Risk	FP16 Conditional Risk	NF4 Refusal	NF4 Joint Risk	NF4 Conditional Risk
0.05	0.842	0.048	0.298	0.854	0.051	0.341
0.10	0.712	0.096	0.331	0.733	0.100	0.371
0.20	0.534	0.199	0.427	0.525	0.202	0.422
0.30	0.377	0.298	0.477	0.364	0.299	0.470
0.40	0.219	0.398	0.509	0.227	0.398	0.515
0.50	0.079	0.498	0.540	0.099	0.498	0.552
0.60	0.002	0.554	0.556	0.003	0.572	0.573

The observed joint risk closely follows the selected CRC target. However, the conditional risk is substantially larger than the corresponding αvalues at the lower targets.

For example, at $\alpha = 0.05$, the conditional risk is 0.298 for FP16 and 0.341 for NF4. This demonstrates that satisfying a joint-risk constraint does not by itself establish the desired conditional-risk guarantee.

C. Bonferroni-Corrected Learn-then-Test

The Bonferroni-corrected Learn-then-Test criteria was considered as the primary conditional-risk criterion. As for stricter cut-off thresholds of 0.05, 0.10, and 0.20, no valid trial values were revealed for both configurations.

At α=0.30, FP16 showed a valid result of solely one applicable trial, whereas NF4 could not demonstrate any applicable trials. In the case of α=0.40, FP16 accomplished 16 valid trials, and NF4 just two valid trials. Such sample sizes are brought under the trajectory of the invalid sample.

At α=0.50, the method became radically better. The number of valid trials for FP16 was 93, which equated to 98.9% of empirical reliability. The number of valid trials for NF4 was 58 with 98.3% reliability.

At α=0.60, both configurations again produced viable results.

TABLE II BONFERRONI-CORRECTED LTT RESULTS

α	*Configuration*	*Refusal*	*Conditional Risk*	*Coverage*	*Reliability*
0.30	FP16	0.961	0.312	0.000	Inconclusive
0.40	FP16	0.772	0.374	0.750	Inconclusive
0.50	FP16	0.645	0.370	0.989	Reliable
0.60	FP16	0.196	0.513	1.000	Reliable
0.40	NF4	0.736	0.467	0.000	Inconclusive
0.50	NF4	0.634	0.421	0.983	Reliable
0.60	NF4	0.252	0.507	1.000	Reliable

The findings suggest the possibility of accurate conditional-risk calibration at about α=0.50+ in the current testing environment. The limit of acceptability was reached at a more relaxed criterion for NF4 than at the stricter thresholds considered earlier.

D. Fixed Sequence Testing

FST produced usable results at several targets where the Bonferroni-corrected procedure was more conservative. For FP16, empirical coverage was 89.1 percent at $\alpha = 0.40$, 94.9 percent at $\alpha = 0.50$, and 100 percent at $\alpha = 0.60$. For NF4, coverage was 0 percent at $\alpha = 0.40$, 79.1 percent at $\alpha = 0.50$, and 93.3 percent at $\alpha = 0.60$. However, the NF4 result at $\alpha = 0.40$ was based on only 4 accepted test examples and is therefore classified as inconclusive under the predefined reliability criterion. In the affected NF4 trials, FST selected thresholds near $\lambda \approx 0.9998$, resulting in only 1–3 accepted test questions. These results, together with the observed lack of consistent monotonicity, motivate treating FST as an exploratory procedure rather than the primary certification method.

E. Effect of Quantization

The distinction between FP16 and NF4 can be noted in terms of the number of occasions on which reliable risk calibration is achieved. If α is set as 0.50, 93 reliable trials can be recorded for FP16 according to the Bonferroni adjustment, while NF4 has produced only 58 such trials.

At the same time the difference can be observed for $\alpha = 0.40$ with FP16 yielding only 16 reliable trials versus 2 for NF4. As long as in both cases the number of reliable trials is below the required threshold of 20 the results obtained using $\alpha = 0.40$ should be qualified as inconclusive ones.

As a result, the findings show clearly that quantization affects the successfulness of achieving reliable conditional calibration in the experiments.

F. Baseline Comparison

The two non-conformal baselines produced substantially different refusal behavior.

TABLE III BASELINE RESULTS

Configuration	Baseline	Refusal	Joint Risk	Conditional Risk
FP16	Prompted refusal	0.125	0.457	0.522
FP16	Fixed threshold 0.5	0.048	0.519	0.546
NF4	Prompted refusal	0.400	0.341	0.568
NF4	Fixed threshold 0.5	0.078	0.512	0.556

The prompted-refusal baseline caused NF4 to refuse 40.0 percent of questions compared with 12.5 percent for FP16. However, the NF4 conditional risk remained 0.568 compared with 0.522 for FP16.

In addition, some of the refusals generated had an answer letter leading with the refusals. These kinds of refusals were classified as refusals according to the pre-defined parsing rule.

V. DISCUSSION

A. Conditional Risk Versus Joint Risk

The output of the CRC brings out the difference between joint risk control and conditional risk control. While the joint risk was close to the set αvalues, the conditional risk still was much higher even at the stricter criteria. As an illustration, at α=0.05, the joint risk was 0.048 and 0.051 for FP16 and NF4 respectively, while the conditional risks were 0.298 and 0.341.

This is due to the fact that joint risk takes into account the probability of rejecting predictions. Thus, a system could lower its joint risk by rejecting many predictions without having a low error in the accepted predictions. It thus comes out clear from the experimental analysis that the calibration process has to correspond with the type of risk to be controlled.

B. Calibration Feasibility

The Bonferroni-corrected Learn-then-Test approach was able to control the conditional risk in question in those trials that had enough data for proper calibration. However, reliable certification was rather hard for tighter error rates.

Specifically, FP16 gave only 16 useful trials while NF4 gave only 2. Both numbers are smaller than the defined reliable limit of 20 accepted examples. In case of α=0.50, the number of reliable trials rose significantly to 93 for FP16 and 58 for NF4.

It is associated with the fact that there are few high-confidence and correct predictions. In case the calibration process chooses a really strict threshold, there will be very little examples to estimate the conditional risk. In such a case, the concentration inequalities will become too conservative to certify the candidate threshold.

Therefore, the lack of ability to certify those targets does not mean that the model is not able to produce them by some possible design of calibration process. Instead, it means that they could not be reliably certified for the model in question, its confidence score, the dataset and the calibration size.

C. Effect of Quantization

The contrast between FP16 and NF4 demonstrates that quantization impacted the viability of conditional-risk calibration.

Reliable Bonferroni-LTT calibration could be performed in 93 percent of cases of FP16 and in 58 percent of cases of NF4 at α=0.50. At α=0.40, the respective numbers were 16 and 2.

It is thus apparent that the impact of quantization is not restricted to memory usage or prediction performance. In the case of this experiment, it additionally manifested itself through variations in the reliability of confidence estimates used for selective predictions.

It should be noted, however, that the above is valid only for Qwen2.5-1.5B-Instruct and NF4 implementation. The current experiment does not prove that all quantization techniques and small language models behave in a similar manner.

D. Limitations of Fixed Sequence Testing

FST yielded good performance in a few cases, but its performance in other cases, especially NF4, was somewhat erratic. Bootstrap showed that a perfectly monotonic risk curve exists in only about 14% to 18% of resamples. Removal of the sparsely populated high confidence tail made the curve more monotonic in FP16 but resulted in a small improvement in NF4.

This indicates that the monotonicity condition assumed by the FST procedure cannot be guaranteed in all cases using confidence scores under investigation. Hence, the FST procedure is not a good substitute for the Bonferroni-corrected procedure.

E. Confidence Quality

The initial technique of using likelihoods for assessing confidence did not help in correctly indicating correctness. The correlation between the technique and the number of answer alternatives was -0.93 while with respect to correctness it was 0.03. The new method of deriving comparative scores based on a letter has improved the correlation of scores with correctness to 0.30 when using FP16 and 0.32 when using NF4 methods. However, dependence on the number of alternatives remained. This is important because conformal calibration cannot indefinitely

compensate for a totally useless or systematically biased confidence score.

F. Prompted Refusal Behavior

The prompted-refusal baseline revealed an important difference between FP16 and NF4. NF4 rejected 40.0% of prompts, whereas FP16 rejected only 12.5%. In spite of the higher refusal rate, NF4 had a conditional risk of 0.568, which is higher than the conditional risk of 0.522 for FP16. This suggests that a refusal decision by the model cannot be considered as reliable uncertainty information. Specifically, in the case-under-review more refusal did not mean lower conditional error.

G. Overall Interpretation

In general, the experiments support three observations. The first observation is that the choice of risk definition plays a critical role since CRC is capable of controlling joint risk without ensuring the required conditional risk. The second one is that calibration is able to ensure conditional risk control, and the reliability of the certification process is highly dependent on the available calibration dataset and the chosen error target. Finally, quantization was found to be related to lower calibration viability and changed confidence behavior in the tested NF4 environment.

H. Limitations

Several limitations exist in this study that should be kept in mind when analyzing the outcomes.

To begin with, there was no reliable certification found for the tighter conditional risk targets below $\alpha=0.30$ with the help of the main Bonferroni-corrected Learn-then-Test approach. Thus, the findings from the experiments do not prove whether the error targets mentioned above are possible to reach using a more substantial calibration set or any other confidence estimator.

The second limitation is associated with the fact that the FP16 results for the target $\alpha=0.40$ included only 16 trials which means that it does not meet the required reliability criterion of 20.

Thirdly, the quantized model utilizes bitsandbytes NF4 instead of GGUF INT4 quantization.

Fourthly, this assessment makes use of one model, which is Qwen2.5-1.5B-Instruct, and one benchmark, which is TruthfulQA MC1. This means that the relationship between the content of the question, the nature of the confidence produced, and the nature of the conformal calibration has not yet been established with regards to a wide range of other models and datasets.

Lastly, the letter-based confidence score has a small correlation with the number of answer options, which is close to -0.23. In addition to being much better than the initial likelihood-based score, this essentially shows that this method has not yet been completely freed from the structure of the questions.

VI. FUTURE WORK

A. Future Work

1. More calibration examples: More calibration examples might make it easier for the concentration-based methods to prove certification for smaller conditional risks.
2. Better confidence predictors: Confidence and nonconformity scores different from the existing ones would diminish the reliance on the number of answer options and increase the gap between correct and incorrect prediction scores.
3. Other quantization schemes: In the current research, we have used NF4 quantization. Analysis of other 4-bit schemes will help to understand whether the behavior of the model was caused by NF4 itself or is rather a general tendency for quantization.
4. Different models and datasets: Analyzing other small language models and datasets would give us information about generalizability of the results obtained in the current study to Qwen2.5-1.5B-Instruct and TruthfulQA.
5. Better monotonicity analysis: More research into conditions of monotonicity of confidence-threshold risk curves will let us understand when conservative approaches like FST can be omitted.

VII. CONCLUSION

This study examines conformal rejection on a small quantized language model working under resource restrictions. The performance of Qwen2.5-1.5B-Instruct was examined on the settings of FP16 and 4-bit NF4 through the tests of TruthfulQA MC1 using confidence-based conformal rejection.

The results indicated that the difference between joint and conditional risks is vital. The CRC method successfully managed the joint risk but failed in

providing the necessary conditional risk control. The Learn-then-Test approach corrected via Bonferroni had a consistent empirical coverage at the loose targets, especially at α=0.50 and α=0.60, although the tight targets were hard to certify.

Furthermore, the comparison between FP16 and NF4 revealed differences in the possibility of calibration. For α=0.50target, successful calibration was achieved in 93% of FP16 runs and 58% of NF4 runs. The findings demonstrate that quantization may influence not only the resource consumption but also the reliability of confidence-based selective prediction in the considered scenario.

In summary, this work proves that the feasibility of conformal refusal can be provided for a small language model in terms of local inference with limited resources. Also, this research revealed the calibration data necessity, confidence quality, and quantization as key factors influencing the conditional risk control reliability. Further research on models, datasets, and quantization approaches is needed to find their applicability.

REFERENCES

[1] C. V. Nguyen *et al*., "A survey of small language models," *arXiv preprint* arXiv:2410.20011, 2024.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[3] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar, "A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions," *ACM Computing Surveys*, vol. 58, no. 3, Art. no. 63, 2025, doi: 10.1145/3744238.

[4] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4878–4887, 2017.

[5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1321–1330, 2017.

[6] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022, doi: 10.18653/v1/2022.acl-long.229.

[7] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proceedings of the International Conference on Learning Representations*, 2024.

[8] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *arXiv preprint* arXiv:2110.01052, 2021.

[9] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," *Journal of the ACM*, vol. 68, no. 6, pp. 1–34, 2021.

[10] Y. Abbasi-Yadkori *et al*., "Mitigating LLM hallucinations via conformal abstention," *arXiv preprint* arXiv:2405.01563, 2024.

[11] B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, and A. Beam, "Conformal prediction with large language models for multi-choice question answering," *arXiv preprint* arXiv:2305.18404, 2023.

[12] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay, "Conformal language modeling," in *Proceedings of the International Conference on Learning Representations*, 2024.