TRAC: Trustworthy Retrieval Augmented Chatbot

Shuo Li [1] **Sangdon Park** [2] **Insup Lee** [1] **Obsert Bastani** [1]

Abstract

Although conversational AIs have demonstrated fantastic performance, they often generate incorrect information, or *hallucinations*. Retrieval augmented generation has emerged as a promising solution to reduce these hallucinations. However, these techniques still cannot guarantee correctness. Focusing on question answering, we propose a framework that can provide statistical guarantees for the retrieval augmented question answering system by combining conformal prediction and global testing. In addition, we use Bayesian optimization to choose hyperparameters of the global test to maximize the performance of the system. Our empirical results on the Natural Questions dataset demonstrate that our method can provide the desired coverage guarantee while minimizing the average prediction set size.

1. Introduction

Neural conversational AIs have recently demonstrated fantastic performance. These chatbots are empowered by large language models (LLMs), and interact with users to perform a number of tasks; we focus on question answering. Although their answers are highly accurate, a major limitation is that these chatbots often confidently generate incorrect responses, called *hallucinations*. Retrieval augmented generation (RAG) has emerged as a promising solution (Lewis et al., 2021; Karpukhin et al., 2020). Given a prompt, these techniques retrieve related contexts that can provide chatbots with helpful information to generate more accurate answers. Also, these techniques can provide timely information by using an up-to-date knowledge base.

In this paper, we explore whether we can provide statistical guarantees for retrieval-augmented question answering systems, to guarantee the system is trustworthy. In particular, we build on *conformal prediction* (Shafer & Vovk, 2007; Vovk & Wang, 2019), a set of tools that modify models to predict sets of labels rather than individual labels. They typically guarantee that the set *covers* the ground truth label with high probability. By predicting a set of labels and providing a coverage, the user can conservatively account for uncertainty in the predicted answer. We propose a novel framework for using conformal prediction to build retrieval augmented question answering systems with high-probability coverage guarantees.

There are several challenges to applying conformal prediction to question answering. First, conformal prediction is usually applied to classification and regression tasks, which are simpler than question-answering. Second, retrieval augmented systems have multiple components, which need to be composed together to form the final prediction. Thus, we need to compose the coverage guarantees of each component to obtain a guarantee for the overall system. Third, conformal prediction typically optimizes a performance metric such as the expected size of the predict label sets, subject to the coverage guarantee. We need to devise reasonable metrics for quantifying the set size for question answering. To our best knowledge, our work is the first to apply conformal prediction to retrieval augmented question answering.

Our framework first construct conformal predictors for the retrieval model and the question answering model, and then combines these techniques by using a multiple hypothesis test (specifically, a *global test*). Given a prompt, the retriever retrieves a set of contexts guaranteed to include the most relevant context with high probability. Then, given the prompt and most relevant context, the LLM predicts a set of answers guaranteed to include the correct answer with high probability. We consider several metrics for evaluating the performance of the final prediction set over answers, including the number of generated answers, the number of unique answers deduplicated by exact match, and the number of generated answers deduplicated by semantic match.

In addition, a key challenge with global tests is that they have hyperparameters that need to be tuned to maximize performance. We propose to use Bayesian optimization to optimize these hyperparameters based on a separate held-out *optimization set*; then, we construct the conformal predictors on a held-out calibration set as usual.

We evaluate our approach on the Natural Question dataset (Kwiatkowski et al., 2019). Our empirical results demonstrate that our approach can provide the desired coverage guarantee, while minimizing prediction set size.

[*]Equal contribution [1]University of Pennsylvania, Pennsylvania, US [2]Georgia Institute of Technology, Georgia, US. Correspondence to: Osbert Bastani <obastani@seas.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Related Work

Retrieval Augmentation. Augmenting chatbots with knowledge from a corpus has shown great effectiveness in reducing hallucinations. Some work focuses on retrieving relevant contexts, such as Karpukhin et al. (2020); Borgeaud et al. (2022). This line of work usually trains a neural retriever to identify relevant contexts for a given question from a knowledge base such as Wikipedia. Other approaches combine training the retriever and the question answerer, including RAG (Lewis et al., 2021) and Atlas (Izacard et al., 2022). Furthermore, instead of retrieving context from an external knowledge base, Wang et al. (2023); Sun et al. (2023) propose to retrieve contexts from another LLM, which is referred to as *parametric memory*. Guu et al. (2020); Lazaridou et al. (2022) focus on designing better in-context prompts so chatbots learn when and what knowledge to retrieve. While these approaches can reduce hallucinations, they do not provide theoretical guarantees.

Conformal Prediction. Conformal prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2007; Angelopoulos & Bates, 2022) is an effective distribution-free uncertainty quantification technique for providing performance guarantees on machine learning models. These techniques construct prediction sets that guarantee to contain true labels with high probability. Split conformal prediction (SCP) (or inductive conformal prediction) reduces the computation complexity of CP by introducing a hold-out calibration set, but maintains the same performance guarantee. CP has been widely applied to image classification (Park et al., 2020; Angelopoulos et al., 2022a; Bates et al., 2021), regression (Lei et al., 2017), object detection (Angelopoulos et al., 2022b).

Global testing. Global testing is a multiple hypothesis testing technique that tests a *global null hypothesis* that consists of all individual hypotheses. Typical tests include the Bonferroni Correction (Bonferroni, 1936), Fisher's Test (Fisher, 1992), and the Harmonic Mean p-value (Wilson, 2019b). Some work has proposed to combine conformal prediction and global testing (Vovk & Wang, 2019; Toccaceli & Gammerman, 2019; Spjuth et al., 2019; Gauraha & Spjuth, 2021; Linusson et al., 2017; Balasubramanian et al., 2014; Toccaceli, 2019). However, these approaches have not been applied to question-answering task; furthermore, they do not use optimization process to improve performance.

3. Methods

3.1. Individual Prediction Sets

First, we describe how we construct prediction sets for the retrieval and question answering models separately using split conformal prediction (SCP). In general, SCP assumes given a *nonconformity measure* $s : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ mapping example-label pairs to scores (typically a pretrained model for solving the task), a held-out calibration set $B = \{(x_i, y_i)\}_{i=1}^N$ sampled i.i.d. from the data distribution $\mathcal{D}$, and a user-specified error level α, and constructs the prediction set $C : \mathcal{X} \to 2^{\mathcal{Y}}$ such that for a new test example x_{N+1} as

$$C(x_{N+1}) = \{y \in \mathcal{Y} \mid s(x_{N+1}, y) \leq \tau\},$$

where τ is the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$-th smallest score in $\{s(x_i, y_i)\}_{i=1}^N$. It guarantees coverage as follows:

Theorem 3.1. *We have*

$$\Pr_{B \sim \mathcal{D}^N, (x,y) \sim \mathcal{D}}(y \in C(x)) \geq 1 - \alpha.$$

Here, the constructed prediction set C implicitly depends on the random calibration set B. In other words, the prediction set $C(x)$ contains the ground truth label for x with probability at least $1 - \alpha$.

To apply conformal prediction to the retrieval and question answering models, the main challenge is to design appropriate nonconformity measures (NCMs) for each task. NCMs are functions measuring how unlikely a given label y is the true label of the observation x. For example, in a multi-classification task, let ξ_k be the estimated for label k, the NCM could be $1 - \xi_k$ for class k.

For the retrieval model, we use the negative inner product or negative cosine similarity between the prompt and context embeddings as the NCM; in both cases, a lower score indicates a higher similarity.

For the question answering model, the NCM is more challenging to design. One option would be the log probability of the generated answer; however, semantically similar answers may induce different log probabilities. To address this limitation, we build on an idea proposed in Kuhn et al. (2023), and propose to use negative *semantic confidence* as the NCM, which we can estimate via Monte Carlo sampling and clustering. In particular, we first request K answers $\{y_k\}_{k=1}^K$; then, we semantically cluster them using an entailment model or their rouge scores; finally, we regard each cluster z as a semantic meaning and estimate its NCM by

$$s_{\text{QA}}(x, y_m; c^*) = -\frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(y_k \in z(x, y_m)),$$

where c^* is the most relevant context for example x. A lower score $s_{\text{QA}}(x, y_m; c^*)$ value indicates that the model is more confident in the semantic meaning of cluster $z(x, y_m)$.

Next, we need to define what is the "ground truth" label, so we can compute the NCM of the true label, which we call the *true label NCM*. For retrieval, consider the true label to be the most relevant context, which is given in the Natural Question dataset. For question answering, given questions and their corresponding top-1 most relevant context, we use

rouge F1 scores (Lin, 2004) along with a standard threshold to determine whether two answers are semantically equivalent. Then, answers that are semantically equivalent to the answer in the dataset are considered ground truth labels.

Now, to construct prediction sets, we split all collected questions into the calibration and testing sets with equal sizes. First, we compute thresholds τ_{ret} as the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$ quantile of the NCMs $s_{\text{ret}}(x, c^*)$ of the true context c^* for question x; and τ_{QA} as the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$ quantile of the NCMs $s_{\text{QA}}(x, y^*; c^*)$ of the correct answer y^* for question x and true context c^*. Then, given a new question x, we construct the retrieval set C_{ret} by including contexts c whose NCMs $s_{\text{ret}}(x, c)$ are no greater than τ^{ret}—i.e.,

$$C_{\text{ret}}(x) = \{c \mid s_{\text{ret}}(x, c) \leq \tau_{\text{ret}}\}.$$

For the question answering model, given a new question x and its most relevant context c^*, we construct prediction sets by including answers y whose corresponding semantic confidence $s_{\text{QA}}(x, y; c^*)$ is no greater than τ_{QA}, i.e.,

$$C_{\text{QA}}(x; c^*) = \{y \mid s_{\text{QA}}(x, y; c^*) \leq \tau_{\text{QA}}\}.$$

Finally, we have the following standard guarantees:

Theorem 3.2. *For retrieval, given a question x and its most related context c^*, we have*

$$\Pr\left(c^* \in C_{ret}(x)\right) \geq 1 - \alpha. \tag{1}$$

For question answering, give a question x, its most related context c^*, and the true answer y, we have

$$\Pr\left(\exists y' \in C_{QA}(x; c^*) \,.\, y' \sim y\right) \geq 1 - \alpha, \tag{2}$$

where $\sim$ denotes semantic similarity.

Note that the randomness is in both with the calibration set and the newly observed example.

3.2. End-to-End Prediction Sets

Next, we describe how we integrate prediction sets for retrieval and question answering to obtain an overall guarantee. The overall prediction set can be obtained by a straightforward composition of the individual prediction sets:

$$C(x) = \bigcup_{c \in C_{\text{ret}}(x)} C_{\text{QA}}(x; c),$$

i.e., run the question answering prediction set on every context. Intuitively, the most related context c^* is contained in $C_{\text{ret}}(x)$ with high probability, and some answer y' semantically equivalent to the true answer y is contained in $C_{\text{QA}}(x; c^*)$, we have

$$y' \in C_{\text{QA}}(x; c^*) \subseteq C(x),$$

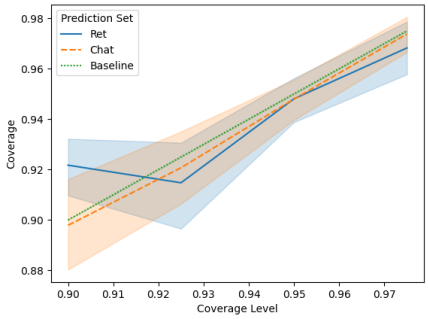

Figure 1. Empirical Coverage Rates with Different Levels

which is the desired guarantee. The main issue is that the individual coverage guarantees only hold with high probability. A naïve strategy is to take a union bound to get

$$\Pr\left(\exists y' \in C(x) \,.\, y' \sim y\right) \geq 1 - 2\alpha. \tag{3}$$

More generally, we can apply global hypothesis tests, which is a technique to efficiently combine multiple statistical tests, to construct prediction sets ((3) corresponds to the Bonferroni correction). Different global tests provide tradeoffs in terms of the resulting assumptions and guarantees. We focus on the Bonferroni Correction (Bonf) and Harmonic Mean p-values (HMP), which are global tests that allow for dependencies between individual tests.

In particular, we treat the individual conformal predictors constructed using split conformal prediction as individual statistical tests, and then combine them with a global test. For Bonf, given error level α, we choose some hyperparameters α_{ret} and α_{QA} such that

$$\alpha_{\text{ret}} + \alpha_{\text{QA}} = \alpha.$$

Then, we compute threshold τ_{ret} using α_{ret} and τ_{QA} using α_{QA} as before. We describe HMP in Appendix C.

Using Bonf, we have the following end-to-end guarantee:

Theorem 3.3. *We have*

$$\Pr\left(\exists y' \in C(x) \,.\, y' \sim y\right) \geq 1 - \alpha.$$

We give a proof in Appendix D. HMP gives a weaker guarantee since it is asymptotic rather than finite sample.

3.3. Hyperparameter Optimization

Many global tests, including Bonf have hyperparameters that need to be tuned. In the Bonferroni Correction, these hyperparameters are α_{ret} and α_{QA}; in HMP, these hyperparameters are weights assigned to different individual tests. Although these hyperparameters do not affect the correctness guarantee, they can significantly affect the resulting prediction set sizes. Thus, to maximize performance, we propose to use Bayesian optimization to choose hyperparameters. We describe our approach in detail in Appendix E.

Table 1. Results on Retrieval Augmented Question Answering. We show coverage rate ("Cov"), # answers ("Ans"), # unique answers by exact match ("Ext") and by Rouge score ("Rou"), and # ChatGPT requests per question ("Req").

		Size			
Method	Cov	Ans	Ext	Rou	Req
CCPS-H	0.91	567.9	21.5	6.1	18.6
CCPS-B	0.90	**523.6**	**20.4**	**5.9**	17.3
HMP	0.92	581.2	23.3	6.5	18.3
Bonf	0.91	530.0	22.3	6.2	**16.7**

We call our method as *Combining Conformal Prediction Sets via Optimized Multiple Hypothesis Testing* as *CCPS*.

Finally, we measure prediction set size in several ways, namely: (i) expected number of answers, (ii) expected number of unique answers deduplicated by exact match, (iii) and expected number of unique answers deduplicated by semantic equivalence based on Rouge score.

4. Experiments

4.1. Experiment Setup

We use Dense Passage Retriever (DPR) as our retriever, and *gpt-3.5-turbo* (ChatGPT) as our question answerer. Our method is agnostic to the retriever and question answerer, and can be straightforwardly adapted to other models.

We evaluate our approach on the Natural Questions dataset (Kwiatkowski et al., 2019). For each question, we retrieve contexts using DPR, and then query ChatGPT on each question-context pair, asking it to return 40 potential answers. One challenge is that querying ChatGPT is costly; thus, we restrict to querying it on the top 20 retrieved contexts per question. We filter out questions for which the most relevant retrieval does not occur in the top 20; this assumption can easily be relaxed by performing additional queries to ChatGPT. While it may increase the overall sizes of the prediction sets, we expect the relative performance of different approaches to be preserved.

We collect 516 examples as the calibration set, 811 data examples as the optimization set, and 812 examples as the test set. We run each experiment with ten random seeds.

We denote CCPS with Bonferroni correction as *CCPS-B* and with Harmonic mean p-value as *CCPS-H*. We compare methods (CCPS-B and CCPS-H) to their counterparts (Bonf and HMP) with $\alpha_{\text{ret}} = \alpha_{\text{QA}} = \frac{\alpha}{2}$.

4.2. Prompt design

To reduce the API cost, we used a prompt that includes both the question and context, but no in-context few-shot demonstrations. To encourage ChatGPT to answer questions based on the retrieved context, we used the following prompt, where <question>is substituted with the question, and <context>with the context:

> Answer the following question based on the given context;
> Answer "I don't know" if you don't know the answer;
> Answer the question using only one keyword.
> Question: <question>
> Context: <context>
> Answer:

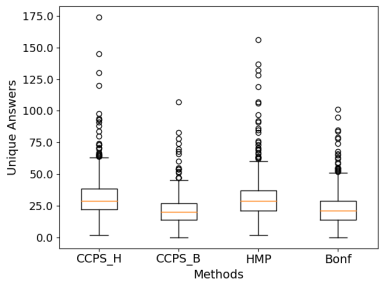

Figure 2. Sizes of Different Methods

4.3. Individual Prediction Sets

We first validate the performance guarantee on retrieval and question answerer prediction sets separately using coverage levels $1-\alpha \in \{0.9, 0.925, 0.95, 0.975\}$. We plot the empirical coverage rate together with the desired coverage level ("Baseline", in dotted green) in Figure 1. As can be seen, the coverage rates are centered around the desired level, which is consistent with (1) & (2).

4.4. End-to-End Guarantee and Performance

We show our results with $\alpha = 0.1$ in Table 1, and results with $\alpha = 0.2$ in Table 2. All approaches satisfied the desired coverage. For number of answers by exact match ("Ext"), CCPS-H and CCPS-B decreased the prediction set sizes of HMP and Bonf by approximately 8%, respectively, demonstrating the benefit of optimization. Furthermore, CCPS-B and Bonf outperformed CCPS-H and HMP, indicating that Bonferroni Correction is more effective for our task. In Figure 2, we plot the distribution of unique answers by exact match (across examples for one random seed). As can be seen, all approaches have similar variation, with CCPS-H and HMP having slightly higher variability. We show some examples of prediction sets in Appendix G.

5. Conclusion

We have proposed a novel strategy applying conformal prediction to retrieval augmented question answering. Our approach ensures an answer semantically equivalent to the true answer is contained in the prediction set, which enables users to act conservatively with respect to the qustion answering system. Our empirical results on the Natural Questions dataset demonstrate that we can obtain coverage guarantees with reasonable prediction set sizes.

References

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction, 2022a.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022b.

Balasubramanian, V. N., Chakraborty, S., and Panchanathan, S. Conformal predictions for information fusion. *Annals of Mathematics and Artificial Intelligence*, 74:45 – 65, 2014.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets, 2021.

Bonferroni, C. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL https://books.google.com/books?id=3CY-HQAACAAJ.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., and Sifre, L. Improving language models by retrieving from trillions of tokens, 2022.

Brown, M. B. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529826.

Fisher, R. A. *Statistical Methods for Research Workers*, pp. 66–70. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_6. URL https://doi.org/10.1007/978-1-4612-4380-9_6.

Gauraha, N. and Spjuth, O. Synergy conformal prediction. In Carlsson, L., Luo, Z., Cherubin, G., and An Nguyen, K. (eds.), *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 91–110. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/gauraha21a.html.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training, 2020.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Atlas: Few-shot learning with retrieval augmented language models, 2022.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih, W. Dense passage retrieval for open-domain question answering, 2020.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lazaridou, A., Gribovskaya, E., Stokowiec, W., and Grigorev, N. Internet-augmented language models through few-shot prompting for open-domain question answering, 2022.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression, 2017.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Linusson, H., Norinder, U., Boström, H., Johansson, U., and Löfström, T. On the calibration of aggregated conformal predictors. In Gammerman, A., Vovk, V., Luo, Z., and Papadopoulos, H. (eds.), *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pp. 154–173. PMLR, 13–16 Jun 2017.

URL https://proceedings.mlr.press/v60/linusson17a.html.

Papadopoulos, H. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Park, S., Bastani, O., Matni, N., and Lee, I. Pac confidence sets for deep neural networks via calibrated prediction, 2020.

Roquain, E. Type i error rate control for testing many hypotheses: a survey with proofs, 2011.

Shafer, G. and Vovk, V. A tutorial on conformal prediction, 2007.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL http://jmlr.org/papers/v9/shafer08a.html.

Spjuth, O., Brännström, R. C., Carlsson, L., and Gauraha, N. Combining prediction intervals on multi-source non-disclosed regression datasets. In Gammerman, A., Vovk, V., Luo, Z., and Smirnov, E. (eds.), *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pp. 53–65. PMLR, 09–11 Sep 2019. URL https://proceedings.mlr.press/v105/spjuth19a.html.

Sun, Z., Wang, X., Tay, Y., Yang, Y., and Zhou, D. Recitation-augmented language models, 2023.

Toccaceli, P. Conformal predictor combination using Neyman–Pearson lemma. In Gammerman, A., Vovk, V., Luo, Z., and Smirnov, E. (eds.), *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pp. 66–88. PMLR, 09–11 Sep 2019. URL https://proceedings.mlr.press/v105/toccaceli19a.html.

Toccaceli, P. and Gammerman, A. Combination of inductive mondrian conformal predictors. *Machine Learning*, 108, 03 2019. doi: 10.1007/s10994-018-5754-9.

Vovk, V. and Wang, R. Combining p-values via averaging, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023.

Wilson, D. The harmonic mean p -value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116:201814092, 01 2019a. doi: 10.1073/pnas.1814092116.

Wilson, D. J. The harmonic mean ¡i¿p¡/i¿-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019b. doi: 10.1073/pnas.1814092116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1814092116.

A. Conformal Prediction and Hypothesis Testing

Conformal prediction is a distribution-free uncertainty quantification technique that constructs provable prediction sets for black-box models. Specifically, let $\mathcal{X}$ and $\mathcal{Y}$ be the input and label spaces, respectively, and (x, y) be an input-label pair. Conformal prediction assumes given a calibration set $B = \{x_n, y_n\}_{n=1}^N$ with N input-label pairs, along with a *nonconformity measure* $s(B, x, y) \in \mathcal{R}$ that measures how different a pair (x, y) is from the examples in B. Given a new input x_{N+1}, conformal prediction constructs a prediction set $C(x_{N+1}) \subseteq \mathcal{Y}$ using Algorithm (Shafer & Vovk, 2008). Intuitively, for every label $y \in \mathcal{Y}$, this algorithm checks whether (x_{N+1}, y) is similar to examples in the B according to the nonconformity measure $s(B, x_{N+1}, y)$. If they are similar, then y is included in the prediction set $C(x_{N+1})$; otherwise, y is excluded from $C(x_{N+1})$. To connect these ideas with multiple hypothesis testing, we note that conformal prediction can be framed as an application of the Neyman-Pearson theory for hypothesis testing (Shafer & Vovk, 2008).

Algorithm 1 The Conformal Algorithm

Input: Nonconformity measure s, significance level α, examples $B = \{x_n, y_n\}_{n=1}^N$, a new input x_{N+1}, label space $\mathcal{Y}$
for $y \in \mathcal{Y}$ **do**
  for $i = 1$ **to** $N + 1$ **do**
    set

$$\alpha_i \coloneqq s(\{(x_1, y_1), \ldots, (x_{N+1}, y)\} \setminus (x_n, y_n), (x_n, y_n))$$

  end for
end for
Set $p_y \coloneqq \frac{\sum_{n=1}^{N+1} \mathbb{I}\{\alpha_i \geq \alpha_{N+1}\}}{N+1}$.
Include y in $C(x_{N+1})$ if and only if $p_y > \alpha$.

A variant of conformal prediction is *inductive conformal prediction (ICP)*, which holds out a fixed calibration set and compares the nonconformity score of new inputs to the calibration set. Since the calibration set is fixed, we omit B in the nonconformity score function. Since we do not need to compute nonconformity scores for the calibration set repeatedly, ICP is more computationally efficient. Papadopoulos (2008) gives a detailed introduction to ICP.

B. Global Testing

Global testing is a technique for combining multiple statistical tests. Individually, each test potentially rejects the null hypothesis, which is referred to as producing a "discovery". The goal of global testing is to minimize false discoveries (i.e., incorrectly rejecting the null hypothesis) by controlling some error rate while maximizing the efficiency of each test (i.e., correctly rejecting the null hypothesis).

Suppose we have a number M of null hypotheses $H^1, \ldots, H^M$. In a single statistical test, we accept the null hypothesis if the test is significant (p-value is sufficiently large) and reject otherwise. After taking these M tests, the possible outcomes are shown in the following table:

	H^m not rejected	H^m rejected	Total
H^m True	$N_{0\|0}$	$N_{1\|0}$	M_0
H^m False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$M - R$	R	M

Here, R is the number of rejections, $N_{0|1}$ and $N_{1|0}$ are the exact (unknown) number of errors made after testing; $N_{1|1}$ and $N_{0|0}$ are the number of correctly rejected and correctly retained null hypotheses. Then, global testing typically controls the *Family-wise error rate* (FWER), which is defined as the probability of falsely rejecting at least one null hypothesis:

$$\text{FWER} = \Pr(N_{1|0} \geq 1).$$

We say an global testing satisfying this bound is *valid*. Common global testing techniques include *Bonferroni correction* (Bonferroni, 1936), Fisher's method (Fisher, 1992), Brown's method (Brown, 1975), and Harmonic mean p-value (Wilson, 2019a).

C. Global Testing via the Harmonic Mean p-Value

HMP is motivated by Bayesian model averaging, and can control of the weak and strong Famility-wise Error Rate (FWER). The control is achieved by combining dependent tests using the generalized central limit theorem.

Specifically, given valid p-values[1] from M statistical tests, denoted as $(p^1, \ldots, p^M)$, and weights for each hypothesis $(w^1, \ldots, w^M)$ satisfying $\sum_{m=1}^{M} w^m = 1$, HMP combines the p-values to form

$$\bar{p} = \frac{\sum_{m=1}^{M} w^m}{\sum_{m=1}^{M} w^m / p^m}. \tag{4}$$

Next, to control the weak FWER at level α, HMP uses the following policy: given the combined p-value $\bar{p}$,

$$\begin{aligned} &\text{If } \bar{p} < \alpha_M: \quad \text{Reject } \{H^1, \ldots, H^M\} \\ &\text{Otherwise:} \qquad \text{Accept } \{H^1, \ldots, H^M\}, \end{aligned} \tag{5}$$

where α_M is an adjusted significance level based on α and the number of test M. In our case ($\alpha = 0.1$, $M = 2$), $\alpha_M = 0.079$. Using this policy, HMP can control of the weak FWER to be under α (Wilson, 2019a). Note that weak FWER equals Type-I error of the global testing when all individual null hypotheses are true (Roquain, 2011).

In the retrieval augmented question answering task, given a question X and a retrieved context c, parameters w_{ret} and w_{QA}, we first compute the p-value for the retrieval task by

$$p_{\text{ret}} = \frac{\sum_{n=1}^{N} \mathbb{I}\left(s_{\text{ret},n} \leq s_{\text{ret}}(X, c)\right)}{N^{\text{ret}}}.$$

We then compare this value to $\lambda = \frac{w_{\text{ret}}}{\frac{1}{0.079} - w_{\text{QA}}}$, which is the minimum p-value for HMP to accept context c. If $p_{\text{ret}} < \lambda$, we reject c; otherwise, we submit the question x together with the context c to ChatGPT, and request answers. Given a generated answer y from ChatGPT, chatbot p-values are computed as

$$p_{\text{QA}} = \frac{\sum_{n=1}^{N} \mathbb{I}\left(s_{\text{QA},n} \leq s_{\text{QA}}(X, y; c)\right)}{N}.$$

Then, Harmonic Mean p-value (HMP) combines these p-values by

$$\bar{p} = \frac{1}{w_{\text{ret}}/p_{\text{ret}} + w_{\text{QA}}/p_{\text{QA}}}.$$

To decide whether to include the answer y, HMP uses the following policy: given the combined p-value $\bar{p}$,

$$\begin{aligned} &\text{If } \bar{p} < \alpha_M: \quad \text{exclude } y \text{ from } C(x) \\ &\text{Otherwise:} \qquad \text{include } y \text{ in } C(x). \end{aligned} \tag{6}$$

D. End-to-End Performance Guarantee

Proof. First, we define the individual null hypothesis in retrieval and chatbot tasks. Given a question X, the null hypothesis for a context ctx is defined as

$$H_c^{\text{ret}} := c \text{ is the most relevant context for } X;$$

given a question X and its top-1 relevant context c^*, the null hypothesis for a generated answer y from the chatbot is defined as

$$H_y^{\text{QA}} := y \text{ is semantically correct for } X \text{ and } c^*.$$

[1] A p-value p for a null hypothesis H^m is valid if it satisfies $\Pr[p \leq \alpha \mid H^m] \leq \alpha$ for all $\alpha \in [0, 1]$, which implies that valid p-values should be subject to a uniform distribution.

Then, we define the global null hypothesis for c and y as the intersection between the two individual hypothesis, i.e.,

$$H_{c,y} = H_c^{\text{ret}} \bigwedge H_y^{\text{QA}},$$

which means that the global null hypothesis is true if both individual hypotheses are true. Using global testing, given a user-specified error level α, we can guarantee that the Type-I error, which is the rate that the true global null hypotheses are rejected, is at most α. In other words, the rate that the true global hypothesis is accepted is at least $1 - \alpha$.

By our null hypothesis definition, a global null hypothesis $H_{c,y}$ is true only if c is the top-1 relevant context and y is a semantically correct meaning. By our algorithm, if the global null hypothesis is true, we will include y into the prediction set. Therefore, the rate that semantically correct meanings ys are included in the prediction set is no less than the rate that the global null hypothesis is accepted, which is at least $1 - \alpha$. Therefore, for the end-to-end prediction set, the semantically correct meanings are included in the set with probability at least $1 - \alpha$, i.e., given a question X and its semantically correct meaning y, we have

$$\Pr\left(y \in C(x)\right) \geq 1 - \alpha.$$

□

Remark D.1. Note that the true meaning coverage rate could be more than the global null hypothesis acceptance rate because semantic meanings based on other relevant contexts could also be correct.

E. Bayesian Optimization

Many global tests have hyperparameters $w \in \mathcal{W}$—e.g., HMP assigns a weight w_{ret} and w_{QA} to each null hypothesis, respectively; and the Bonferroni Correction assigns a significance level α_{ret} and α_{QA} to each hypothesis. While these parameters do not affect the Type-I error rate of the global test, they can affect the Type-II error rate and therefore the resulting cost of $C_{B,w}$.

Our method uses Bayesian Optimization (BO) to optimize these hyperparameters $w \in \mathcal{W}$ to minimize the given cost g. In particular, BO first initializes a Gaussian Process (GP) model of the cost function. Then, based on the GP, BO selects parameters potentially minimizing the cost function and evaluates the prediction set cost on the selected parameters. Finally, BO refines the GP model based on the evaluated cost. BO iteratively optimizes the objective function across T iterations.

To preserve the validity of the global test, we separate global testing from BO. In particular, we split the available data into a calibration set and an optimization set (we also use a separate training set to train the nonconformity scores s^m, but this step occurs prior to applying CCPS). The parameters w are first optimized by running a global test on the optimization set, and evaluating the resulting cost. Once we have chosen hyperparameters w, CCPS runs the global test one final time, but now in conjunction with the held-out calibration set B, to obtain $C_{B,w}$. The pseudo-code can be found in Algorithm 2.

Algorithm 2 Trustworthy Retrieval Augmented Chatbot (TRAC)

Input: global test with hyperparameters $w \in \mathcal{W}$, dataset $B = \{x_n, c_n, y_n\}_{n=1}^{2N}$, desired error rate α,
Split B into optimization set $B_{\text{opt}} = \{(x_n, c_n, y_n)\}_{n=1}^{N}$ and calibration set $B_{\text{cal}} = \{(x_n, c_n, y_n)\}_{n=N+1}^{2N}$
Initialize Gaussian process G
Compute nonconformity scores $s_{\text{ret},n}(x_n, c_n)$ and $s_{\text{QA},n}(x_n, y_n; c_n)$ for each $(x_n, c_n, y_n) \in B$
for $t \in \{1, ..., T\}$ **do**
  Choose hyperparameters $w_t \in \mathcal{W}$ using Bayesian optimization on G
  Construct prediction set C_{B_{opt}, w_t} using the given global test
  Compute the empirical prediction set cost $c_t \leftarrow \frac{1}{K} \sum_{x \in B_{\text{opt}}} g(C_{B_{\text{opt}}, w_t}(x_i))$
  Update G using (w_t, c_t)
end for
Let w^* to be the hyperparameters w_t with the smallest cost c_t (over $t \in \{1, ..., T\}$)
Return prediction set C_{B_{cal}, w^*} constructed using the given global test

F. Results with $\alpha = 0.2$

Table 2. Results on Retrieval Augmented Question Answering. We show coverage rate ("Cov"), # answers ("Ans"), # unique answers by exact match ("Ext") and by Rouge score ("Rou"), and # ChatGPT requests per question ("Req").

		Size			
Method	Cov	Ans	Ext	Rou	Req
CCPS-H	0.84	485.3	16.5	5.0	18.1
CCPS-B	0.83	**409.9**	**14.2**	**4.4**	16.2
HMP	0.85	502.8	17.9	5.3	17.7
Bonf	0.86	428.1	15.9	5.7	**15.4**

G. Examples of Prediction Sets

Quesiton: 'what is the second movie of the pirates of the caribbean'

Reference Answer: "Dead Man 's Chest"

Answer Set: "Dead Man's Chest.", 'Dead Men Tell No Tales', "Don't know.", 'Pitch Black', "Dead Men Tell No Tales or Salazar's Revenge", "I"Don't know"", "don't know", "Unknown/ I don't know.", "Dead Men Tell No Tales/Salazar's Revenge", "Dead Men Tell No Tales (or Salazar's Revenge)", '"On Stranger Tides".', "I Don't Know", 'Unknown.', '"I doné know"', 'fourth.', 'Pirates', 'Pirates.', "Don't know", '"Unknown"', "Don't remember/I don't know", 'fourth', "I don't know", "dead man's chest", 'Dead Men Tell No Tales.', 'On Stranger Tides.', '"fourth"', "I Don't know.", '"Doné know"', 'unknown', '"Dead Mané Chest"', "Dead Man's Chest (keyword: Chest)", "Dead Man's Chest", "Dead Men's Chest", "Dead Man's Chest", '"On Stranger Tides"', '"Pirates"', "I don't know.", 'Unknown', 'Dead Men Tell No Tales (or fifth film)', 'Fourth', '"Fourth"', 'unknown.', 'Pitch Black.', '"Dead Men Tell No Tales"', '"Pirates."', 'On Stranger Tides'

Quesiton: "when did spanish town become jamaica 's capital"

Reference Answer: "1534"

Answer Set: '1680', "Don't know.", '1534.', '1873.', '1655', '1845', "don't know", '1872.', '1534', 'Eighteenth Century', 'Eighteenth century.', '1670.', '1845.', 'Eighteenth century', "Don't know", 'eighteenth century', "I don't know", '1873', '1847', "Not mentioned/ I don't know", "I don't know.", '1670', '1680.', 'eighteenth century.', '1962', '1962.', '1847.', '1872', '1655.'

Quesiton: 'who presented in parliament the separate rail budget in india'

Reference Answer: "'the Minister of Railways'"

Answer Set: 'Lalu Yadav.', 'Minister of Railways.', 'D. V. Sadananda Gowda', "Don't know.", 'Ms. Mamata Banerjee', 'parliament', 'D. V. Sadananda Gowda.', 'Sir William Acworth', 'Lalu Prasad Yadav.', 'John Mathai.', 'John Mathai', '"I doné know"', 'Minister.', 'Minister of Railways', 'Minister', 'Suresh Prabhu', '"I doné know".', "I don't know", 'RLDA', 'Parliament.', 'Sir William Acworth.', 'RLDA.', 'Mamata Banerjee', "I Don't Know.", 'Lalu Yadav', "I don't Know.", 'Lalu Prasad Yadav', 'D.V. Sadananda Gowda.', "I don't know.", 'Suresh Prabhu.', 'Mamata Banerjee.', 'D.V. Sadananda Gowda', 'Parliament'