The Effects of Data Augmentation on Confidence Estimation for LLMs

Anonymous ACL submission

Abstract

Confidence estimation is crucial for reflecting the reliability of large language models (LLMs), particularly in the widely used closed-source models. Utilizing data augmentation for confidence estimation is viable, but discussions focus on specific augmentation techniques, limiting its potential. We study the impact of different data augmentation methods on confidence estimation. Our findings indicate that data augmentation strategies can achieve better performance and mitigate the impact of overconfidence. We investigate the influential factors related to this and discover that, while preserving semantic information, greater data diversity enhances the effectiveness of augmentation. Furthermore, the impact of different augmentation strategies varies across different range of application. Considering parameter transferability and usability, the random combination of augmentations is a promising choice. Our codes and data are available at: https://anonymous.4open.science/r/ceda.

1 Introduction

Although LLMs (Chowdhery et al., 2023; OpenAI, 2023) exhibit remarkable capabilities in generalization across various natural language processing (NLP) tasks, their tendency to generate non-factual responses (Ji et al., 2023) raises significant concerns. It is essential to assess the reliability of their generalization results. As black-box LLMs become more prevalent, accessing their internal information is challenging, which increase the difficulty of evaluate the reliability. Confidence estimation (Tian et al., 2023; Xiong et al., 2024; Zhang et al., 2024) has emerged as a popular solution, facilitating risk assessment and error checking.

There is a wide variety of methods (Lin et al., 2024; Liu et al., 2024; Xie et al., 2024) for confidence in white-box LLMs, but there are few types of methods for black-box LLMs, where the common methods are Bayesian methods (Xiong et al., 2024; Ling et al., 2024), particularly in sampling strategy (Si et al., 2023). It inputs the original text into the LLM and samples for predictions. In addition to that, some researchers (Hashimoto et al., 2024; Xiong et al., 2024; Gao et al., 2024) using data augmentation for confidence estimation. Arguably, it is a promising strategy for LLM's confidence estimation, especially for black-box ones with inaccessible parameters. However, they either operate only on white-box small models or engage in very limited discussions regarding augmentation methods, relying on weak augmentation or uncontrollable augmentation. Therefore, we hope to conduct a comprehensive evaluation of confidence estimation based on data augmentation to verify its impact.

In this study, we explore common augmentation strategies (Wei and Zou, 2019; Sennrich et al., 2016) and discuss typical automated augmentation methods (Cubuk et al., 2020; Ren et al., 2021) to validate the effectiveness of combining augmentation techniques. In particular, we investigate the following research questions:

- **Q1**: What impact does data augmentation have on confidence estimation?
- **Q2**: What factor contributes to this impact?
- **Q3**: Is the range of applications for different data augmentation techniques consistent?

Comprehensive experiments on benchmark datasets show that data augmentation is effective for confidence estimation and mitigates the impact of LLMs' overconfidence. The best data augmentation method reduces the average ECE across three models from 11.50% to 5.97% in GSM8K. We perform further analysis of the experimental results to explore which factors impact the results of data augmentation strategies on confidence estimation. Our findings indicate that data diversity and semantic consistency are key. While maintaining semantic information, higher data diversity leads

to improved confidence estimation. Notably, RandAugment (Cubuk et al., 2020) demonstrates better performance and exhibits significant potential for cross-model transfer. Moreover, our study reveals that different data augmentation methods have different ranges of applicability. A mild augmentation strategy is more appropriate for mathematical data about complex logical reasoning. Therefore, when addressing downstream tasks that are unknown, it is advisable to use RandAugment to compute confidence. We believe this will contribute to enhancing the reliability of generalization in LLMs.

2 Related Works

Confidence Estimation for Black-Box LLMs. The confidence estimation of black-box models is divided into four categories by common methods. Single deterministic methods (Lin et al., 2022; Tian et al., 2023) rely on verbal descriptions, which are hard to align with the model's internals (Kumar et al., 2024). Ensemble methods (Zhang et al., 2023) based on multiple LLMs diverge from our focus on a single model, so we won't explore this. While Bayesian methods (Xiong et al., 2024; Zhang et al., 2024) are common, LLMs' overconfidence undermines their effectiveness. In the methods that use augmentation, the weak augmentation (Gao et al., 2024) struggles to enhance text diversity. Paraphrasing method (Xiong et al., 2024) lacks controllability. This makes it difficult to mitigate overconfidence.

Data Augmentation. Data augmentation (Wei and Zou, 2019; Feng et al., 2021) is widely used in NLP, primarily includes insertion, deletion, replacement, and swapping. Back-translation (Sennrich et al., 2016) is useful in low-resource domains. With the rise of LLMs, paraphrasing method (Piedboeuf and Langlais, 2023) using LLMs rewrite the text, but ensuring quality can be difficult and expensive. Besides, automated augmentation has gained attention. Text AutoAugment (Ren et al., 2021) uses automated augmentation in NLP, while RandAugment (Cubuk et al., 2020) reduces retrieval costs. In subsequent experiments, we examine traditional and automated augmentation, revealing their potential in confidence estimation.

3 Methodology

3.1 Confidence Estimation

We utilize data augmentation strategies to compute confidence. Specifically, given a sample X,

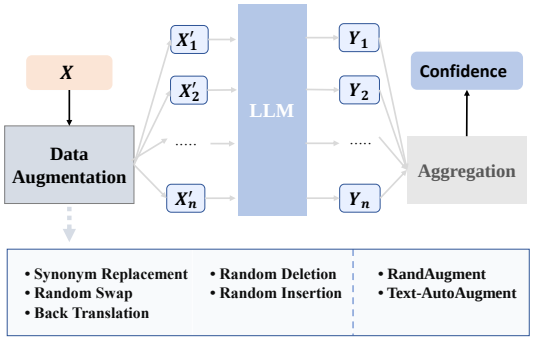

Figure 1: Overview of a confidence estimation framework utilizing data augmentation. Each sample X is subjected to n augmentations, yielding a more diverse array of augmented instances X'. The LLM performs predictions Y on these augmented samples, and confidence is derived through aggregation.

it undergoes augmentation n times to generate n augmented samples $\mathcal{S} = [X'_1, X'_2, ..., X'_n]$, where $X'_i = s_i(X)$ and $s_i(\cdot)$ is the i-th data augmentation for the sample. These augmented samples $\mathcal{S}$ are input into the LLM f to produce predictions $\mathcal{A} = [Y_1, Y_2, ..., Y_n]$, i.e., $Y_i = f(X'_i)$. Subsequently, a consistency strategy (Si et al., 2023) is applied to aggregate the answers. The confidence C about X is the consistency of answer $\bar{Y}$ in $\mathcal{A}$,

$$C = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i = \bar{Y}\}, \tag{1}$$

where $\bar{Y}$ is the most frequently predicted answer and $\mathbb{I}$ is indicator function.

3.2 Data Augmentation Methods

The selection of augmentation strategy is important in our confidence computation. We believe that an effective augmentation strategy should be controllable and simple. Consequently, we choose the straightforward method from the four most common categories of traditional augmentation techniques: **synonym replacement**, **random swap**, **random deletion**, and **random insertion**. Each method enables control over the degree of augmentation by adjusting the magnitude. Besides, we discuss a popular data augmentation technique, **back-translation** (Sennrich et al., 2016).

To examine the synergistic effects of traditional augmentation strategies, we discuss two automated augmentation methods: **RandAugment** (Cubuk et al., 2020) and **Text-AutoAugment (TAA)** (Ren et al., 2021). RandAugment randomly combines traditional augmentation strategies, while TAA applies a fixed set of augmentation to each sample.

Method	StrategyQA			Professional Law			GSM8K			AVG
	Qwen2	Llama3	Gemma2	Qwen2	Llama3	Gemma2	Qwen2	Llama3	Gemma2	
Sampling	27.00	28.25	22.14	42.04	39.32	33.22	12.94	12.14	9.41	25.16
Paraphrase	33.57	35.33	39.42	48.46	46.22	45.54	46.13	61.44	67.46	47.06
Synonym Replacement	23.72	26.01	20.21	38.98	36.41	30.86	**5.38**	8.56	**4.72**	21.65
Random Swap	23.81	29.38	20.76	35.58	36.96	30.39	6.75	10.48	25.09	24.36
Random Deletion	20.46	**20.81**	**16.28**	37.65	36.56	29.73	8.74	7.81	13.04	21.23
Random Insertion	26.66	28.37	20.61	38.41	37.29	32.30	13.44	14.36	11.15	24.73
Back-Translation	24.13	27.94	22.20	46.67	44.59	38.03	8.05	8.36	6.14	25.12
RandAugment	20.55	26.73	16.46	**33.13**	36.85	**26.92**	6.55	**7.17**	6.15	**20.05**
TAA	**18.98**	27.38	20.61	38.17	**34.96**	31.05	7.65	19.27	15.38	23.72

Table 1: Confidence estimation of 3 models (metrics are given by $\times 10^2$). The evaluation metric is ECE(↓). The best results are marked in bold and the second-best marked in underline.

Method	StrategyQA		Professional Law		GSM8K		AVG
	GPT-3.5	GPT-4o-mini	GPT-3.5	GPT-4o-mini	GPT-3.5	GPT-4o-mini	
Sampling	18.62	18.45	30.77	29.36	8.94	3.83	18.33
Synonym Replacement	19.59	16.05	31.92	28.54	7.33	3.81	17.87
Random Deletion	**16.68**	**10.86**	27.86	26.72	**4.84**	**3.75**	**15.12**
RandAugment	18.68	16.73	**25.42**	**22.97**	7.11	7.02	16.32

Table 2: Further experiments about the best method from each category for the closed-source model (metrics are given by $\times 10^2$). The evaluation metric is ECE(↓).

They need to find the best parameters in the validation set and use them for the test set. More details are described in the Appendix.

4 Experiments

4.1 Experimental Setup

To discuss the impact of data augmentation on confidence estimation, we conduct experiments using Llama-3-8b (AI@Meta, 2024), Gemma-2-9b (Rivière et al., 2024), and Qwen-2-7b (Yang et al., 2024). We use GPT-3.5, and GPT-4o-mini for further experimental verification. The implementation details are in the Appendix B.

Datasets. We evaluate the quality of confidence estimates across three datasets: 1) *StrategyQA* (Geva et al., 2021) from BigBench (Ghazal et al., 2013) , which is about commonsense reasoning ; 2) *Professional Law* (Prf-Law) from MMLU (Hendrycks et al., 2021), which is about professional knowledge ; 3) *GSM8K* (Cobbe et al., 2021) is about math word problems.

Baselines. We select representative strategies from two categories. **The baselines utilize Eq.(1) for aggregation**. For the Bayesian method, we use *sampling strategy* (Si et al., 2023) that inputs the original text into the LLM, with responses sampled at a high temperature. Besides, we use a *paraphrasing-based method*, which paraphrases questions using LLMs (Xiong et al., 2024).

Evaluation Metric. To assess the alignment between confidence and accuracy, we introduce Expected Calibration Error (ECE) (Naeini et al., 2015). It calculates the difference between the average confidence and the accuracy of the model.

4.2 Results

Answer for Q1: Data augmentation is beneficial for confidence estimation, which mitigates the impact of LLMs' overconfidence. As shown in Table 1, the average ECE of all augmentation strategies outperforms the sampling method. The top three performing augmentation strategies are *RandAugment, Random Deletion, and Synonym Replacement*. In GSM8K, Synonym Replacement reduces the average ECE across three models from 11.50% to 5.97%. Additionally, there are decreases of 6.61% and 5.89% in StrategyQA and Professional Law, respectively. In Table 2, the average ECE of data augmentation methods still demonstrates performance advantages.

The hallucinations of LLMs often lead to overconfidence in incorrect predictions. We compare the confidence of three optimal augmentation strategies on incorrect samples to further investigate the

Method	StrategyQA	Prf-Law	GSM8K
Sampling	0.89	0.84	0.44
Synonym Replacement	0.85	0.79	0.36
Random Deletion	0.79	0.76	0.35
RandAugment	0.79	0.72	0.34

Table 3: Mean confidence for Qwen2, Llama3, and Gemma2. The table presents the average confidence (↓) of incorrectly predicted samples.

Model	Method	StrategyQA	Prf-Law	GSM8K
Llama3	Sampling	28.25	39.32	12.14
	RandAugment	26.73	36.85	7.17
	+ Parameters	26.73	36.27	15.61
Gemma2	Sampling	22.14	33.22	9.41
	RandAugment	16.46	26.92	6.15
	+ Parameters	19.62	24.97	9.33

Table 4: Apply the RandAugment parameters from the smallest model, Qwen2-7b, to the larger models, Llama-3-8b and Gemma-2-9b, as indicated in grey. The evaluation metric is ECE(↓). Metrics are given by $\times 10^2$.

impact of data augmentation. Ideally, the confidence for incorrect samples should be 0. However, as shown in Table 3, the sampling method exhibits high confidence, particularly in StrategyQA and Professional Law, where the confidence reaches up to 0.89 and 0.84. Based on data augmentation strategies, the overconfidence of LLMs is alleviated, and the confidence is reduced to 0.79 and 0.72. It shows that data augmentation holds significant potential for confidence estimation.

Answer for Q2: Data diversity and semantic consistency are important. While maintaining semantic integrity, the more diverse the samples are, the better the outcomes will be. Based on the analysis of Table 1, the degradation of semantic information, such as random swap, results in poorer outcomes. Similarly, random insertion introduces noisy information that can affect the judgments of LLMs. Moreover, over-focusing on semantic consistency and ignoring sample diversity can also be detrimental to confidence estimation, as seen with back-translation techniques. We speculate that the effectiveness of random deletion is from the LLM's inherent ability to infer missing tokens, allowing it to extract meaningful information from diverse augmented samples. RandAugment ranks highest in average ECE in Table 1 because its random combinations of individual strategies enhance sample diversity. In contrast, the fixed augmentation combinations used in TAA result in lower diversity compared to RandAugment, which is why TAA performs worse than RandAugment.

We conduct a further analysis of RandAugment and find that it demonstrates **cross-model adaptability**. Specifically, the augmentation combinations learned by a model can be transferred to other models. In Tables 4, we apply the best parameters from Qwen2 to Llama3 and Gemma2, resulting in performance improvements in most cases compared to the sampling method. In StrategyQA, the best parameters from Llama3 and Qwen2 are the same, so they achieve the same results.

Answer for Q3: Different data augmentation techniques have different ranges of applications. Moderate strategies are recommended only for math data that requires complex reasoning. In Table 1 and Table 2, RandAugment and random deletion, which can introduce more diverse samples, generally achieve favorable outcomes across a range of datasets. However, relatively mild augmentation strategies often demonstrate significant dataset-specific biases. Back-translation perform better on GSM8K compared to StrategyQA and Professional Law. While back-translation results underperforms on Professional Law, it reduces the ECE on GSM8K from 11.50% to 7.52%. For a math dataset that prioritizes logical reasoning, the complexity of the questions and the requirement for logical reasoning make the dataset more sensitive to sample diversity. Thus, milder augmentation strategies remain effective. However, for StrategyQA and Professional Law, the strong common-sense reasoning and extensive domain knowledge in LLMs render overly cautious augmentation strategies ineffective.

5 Discussion

LLMs are sensitive to prompts (Sclar et al., 2024), making responses from diverse views align better with their cognition, leading to the effectiveness of data augmentation in confidence estimation. Given the different applicability of augmentation strategies, we recommend using RandAugment for confidence calculation in unknown downstream tasks. This is due to its cross-model adaptability and usability. We are confident that the confidence estimation based on data augmentation will beneficial for the reliability of LLM generalization, supporting the development of safer and reliable NLP systems.

Limitation

While we conduct experiments and validate the effectiveness of data augmentation in confidence estimation, there is a limitation that should be acknowledged. Regarding the data augmentation strategies, due to resource limitations, we don't discuss all possible augmentation techniques; instead, we only consider five typical traditional augmentation strategies and two basic automatic augmentation strategies. We acknowledge the possibility that some augmentation methods that we don't discuss may be more suitable in certain contexts than others that we have considered.

Ethical Considerations

In this study, we utilized existing datasets that have already addressed ethical considerations. Additionally, the data augmentation methods employed are safe and reliable, making it unlikely that toxic sentences will be generated. This has been validated by numerous previous studies. Additionally, our manual reviews did not reveal any issues.

References

AI@Meta. 2024. Llama 3 model card.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3008–3017. Computer Vision Foundation / IEEE.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2336–2346. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: towards an industry standard benchmark for big data analytics. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 1197–1208. ACM.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. 2024. Are data augmentation methods in named entity recognition applicable for uncertainty estimation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18852–18867. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances*

in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 315–334. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10351–10368. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3357–3370. Association for Computational Linguistics.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. 2024. Enhancing language model factuality via activation-based confidence calibration and guided decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10436–10448. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press.

OpenAI. 2023. Gpt-4 technical report.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is chatgpt the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15606–15615. Association for Computational Linguistics.

Shuhuai Ren, Jinchao Zhang, Lei Li, Xu Sun, and Jie Zhou. 2021. Text AutoAugment: Learning compositional augmentation policy for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18128–18138. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sricharan. 2023. Sac^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15445–15458. Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2959–2979. Association for Computational Linguistics.

A Additional Experimental Results

Transferability across datasets. We discuss the adaptability of RandAugment across different datasets. It is evident that, in most cases, the parameters obtained from other datasets maintain relatively stable performance on the target dataset. This adaptability is particularly noteworthy when the target dataset is StrategyQA or Professional Law, demonstrating strong cross-dataset compatibility.

Target Dataset	Source Dataset	Qwen2	Llama3	Gemma2
StrategyQA	StrategyQA	20.546	26.726	16.463
	Prf-Law	20.813	27.467	18.713
	GSM8K	21.564	26.167	16.463
Prf-Law	StrategyQA	32.840	37.001	24.424
	Prf-Law	33.129	36.852	26.919
	GSM8K	36.587	34.713	24.424
GSM8K	StrategyQA	16.61	10.571	6.149
	Prf-Law	9.314	17.750	32.724
	GSM8K	6.547	7.170	6.149

Table 5: Apply the parameters of source dataset to target dataset.

B Implementation Details

We follow prior work (Xiong et al., 2024) to set the augmentation times to $n = 5$ and keep the sampling quantity of the baseline consistent with it. The model temperature is 0 for the paraphrasing methods, and 0.7 for the sampling-based and our method. The validation set to test set ratio is 1:1. Augmentation magnitude can be selected from $\{0.1, 0.2, 0.3\}$. Automated augmentation retrieval process assesses the quality of augmentation combinations based on ECE, excluding model training. For all experiments, we run three times and report the averaged results.

B.1 Traditional Data Augmentation

We provide supplementary information about traditional data augmentation:

- **Synonym Replacement (SR).** SR (Wei and Zou, 2019) randomly selects words from the sentence. It then substitutes each of these words with a randomly chosen synonym.

Dataset	Prompt
StrategyQA & Prf-Law	Read the question, analyze step by step, provide your answer. Use the following format to answer: "'Explanation: [insert step-by-step analysis here] Answer: [ONLY the option letter; not a complete sentence]"' Only give me the reply according to this format, don't give me any other words.
GSM8K	Read the question, analyze step by step, provide your answer. Use the following format to answer: "'Explanation: [insert step-by-step analysis here] Answer: [ONLY the number; not a complete sentence]"' Only give me the reply according to this format, don't give me any other words.

Table 6: Prompt templates for each dataset.

- **Random Swap (RS).** RS (Wei and Zou, 2019) randomly swaps the positions of two words K times.

- **Random Deletion (RD).** RD (Wei and Zou, 2019) randomly deletes each word based on a probability of p.

- **Random Insertion (RI).** RI (Wei and Zou, 2019) finds a random synonym of a random word in the sentence that is not a stop word and inserts that synonym into a random position in the sentence K times.

- **Back-Translation.** It (Sennrich et al., 2016) typically translates text from the original language to a second language, and then back to the original language. Here, we choose French as the second language.

In the experiment, aside from back-translation, which does not require an augmentation magnitude, the remaining four augmentation strategies that need to be tested on the validation set should determine the optimal augmentation magnitude before being applied to the test set.

B.2 Automated Augmentation

We describe the details of the automated augmentation strategies:

- **RandAugment.** It is necessary to determine the optimal augmentation combination on the validation set before applying it to the test set. The key parameters are the number of augmentation transformations N_r and the magnitude of augmentation M. For each sample, N_r augmentation transformations are randomly selected from five traditional augmentation operations and then applied sequentially to the sample, each applied with a magnitude of M. The range of values for the number of augmentation transformations N_r is {1, 2, 3}. The magnitude M can take values from {0.1, 0.2, 0.3}. A grid search should first be conducted on the validation set to determine the optimal values for N_r and M, which will then be applied to the test set.

- **TAA.** It also requires determining the optimal augmentation combination first. A augmentation combination consists of N_t policies, with each policy $P = \{O_1, O_2, ..., O_T\}$ containing T editing operations O. Each editing operation includes an augmentation transformation, a probability of calling a transformation, and a magnitude. Each sample requires executing all editing operations in a policy, i.e., $s = O_T \circ ... \circ O_2 \circ O_1$.

In TAA, the number of policies is $N_t = 4$, and the number of editing operations is $T = 2$, with the top three optimal combinations retained, i.e., $3 * N_t$. We set the number of iterations to 50.

B.3 Additional Implementation Details

To ensure that the LLM thoroughly understands the questions and accurately demonstrates its capabilities while minimizing the occurrence of hallucinations, we employ zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022).