SHORT-PAPER

Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models

PRASENJIT DEY, Amazon.com, Inc., Seattle, WA, United States

SRUJANA MERUGU, Amazon.com, Inc., Seattle, WA, United States

SIVARAMAKRISHNAN KAVERI, Amazon.com, Inc., Seattle, WA, United States

Published: 08 May 2025

WWW '25: The ACM Web Conference 2025
April 28 - May 2, 2025
Sydney NSW, Australia

Conference Sponsors:
SIGWEB

WWW '25: Companion Proceedings of the ACM on Web Conference 2025 (May 2025)
https://doi.org/10.1145/3701716.3715523
ISBN: 9798400713316

Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models

Prasenjit Dey
Amazon
Bengaluru, India
prasendx@amazon.com

Srujana Merugu
Amazon
Bengaluru, India
smerugu@amazon.com

Sivaramakrishnan Kaveri
Amazon
Bengaluru, India
kavers@amazon.com

Abstract

Large Language Models (LLMs) are known to hallucinate and generate non-factual outputs which can undermine user trust. Traditional methods to directly mitigate hallucinations, such as representation editing and contrastive decoding, often require additional training data and involve high implementation complexity. While ensemble-based approaches harness multiple LLMs to tap into the "wisdom of crowds", these methods overlook uncertainties in individual model responses. Recent studies reveal that uncertainty estimation can enable LLMs to self-assess the likelihood of generating hallucinations. In this work, we focus on factoid question answering (QA) and observe that LLMs accuracy and self-assessment capabilities vary widely with different models excelling in different scenarios. Leveraging this insight, we propose Uncertainty-Aware Fusion (UAF), an ensemble framework to reduces hallucinations by strategically combining multiple LLM based on their accuracy and self-assessment abilities. Empirical results on several public benchmark datasets show that UAF outperforms state-of-the-art hallucination mitigation methods by 8% in factual accuracy, while either narrowing or surpassing the performance gap with GPT-4.

CCS Concepts

• **Computing methodologies → Ensemble methods**; **Natural language generation**.

Keywords

Large Language Models, Hallucination detection, Uncertainty, Ensemble

ACM Reference Format:
Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. 2025. Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715523

1 Introduction

Large language models (LLMs) have yielded remarkable performance boosts across diverse natural language processing (NLP) tasks [1]. However, their tendency to 'hallucinate', i.e., generating outputs not grounded in factual training data[6], remains a critical limitation, impeding deployment, particularly in high-stakes applications where factual accuracy is paramount.

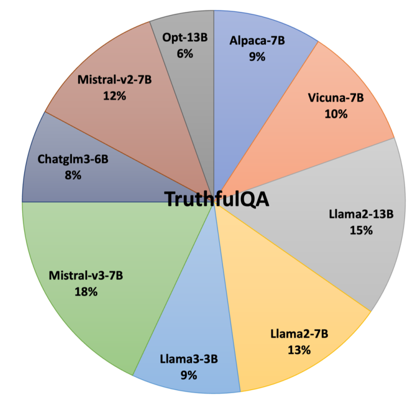

Figure 1: Fraction of examples where each LLM generates the most confident correct answer. Optimal LLM (high accuracy and confidence) varies across examples.

Related work: Existing work on hallucination mitigation can be broadly categorized into three groups. The first group of techniques tries to enhance truthfulness at inference time using contrastive decoding or representation editing. Representation editing methods like ITI [12] and TruthX [21] identify a truthful direction using human-labeled data and adjust model activations accordingly. Contrastive decoding approaches such as SH2 [10] and DoLa [3] modify output probabilities by comparing distributions between different models or layers. However, these methods often face challenges with implementation complexity, reliance on extensive labeled data, and limited generalization across domains.

The second group employs ensemble learning techniques to improve model performance. Methods such as LLM-Blender [7] learn to combine diverse candidate responses from multiple LLMs, while MLM [18] selects the response most similar to the input. Another approach exemplified by Consistent [19], generates multiple candidate responses and uses majority voting for the final output. However, these approaches often overlook the inherent uncertainties in candidate responses, potentially limiting their effectiveness.

The third group of methods is based on the observation that LLMs often have an internal sense of truthfulness, even when producing false statements [2, 9]. This insight is exploited in various approaches such as logit probability-based methods [14, 17], consistency-based techniques [11, 15], probing internal representations [2, 4], and prompting-based strategies [9] to obtain uncertainty scores without additional training or model editing, offering promising

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715523

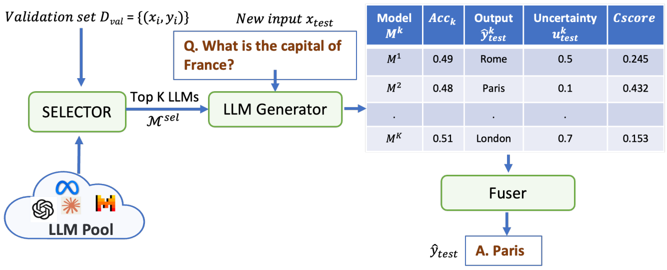

Figure 2: Overview of our UAF architecture

avenues for hallucination detection and mitigation.

Contributions: In this work, we focus on mitigating LLM hallucinations for factoid question answering (QA). First, we observe that the proliferation of open-source LLMs, each with distinct strengths and weaknesses, precludes any single model from consistently outperforming others in factual accuracy and self-assessment. Figure 1 illustrates this phenomenon for the TruthfulQA [13] dataset, showing the distribution of examples where each LLM produces the most confident correct response. Despite Mistralv3-7B achieving the highest overall accuracy, it leads in confidence for only 18% of examples, indicating that the rest are better handled by other models. This diversity in model performance across examples presents an opportunity to leverage the complementary strengths of multiple LLMs to reduce hallucinations and enhance overall accuracy. Our key contributions are summarized below:

1. We demonstrate, through extensive experiments with six LLMs across multiple benchmarks, that the relative factual accuracy and self-assessment capabilities of LLMs vary significantly across examples, with no single dominant model.
2. We propose Uncertainty-Aware Fusion (UAF), an ensemble framework that strategically combines multiple LLMs to achieve superior performance. UAF consists of two key modules: (i) SELECTOR, which chooses the top K LLMs from the entire pool based on accuracy and hallucination detection abilities, and (ii) FUSER, which merges the outputs of these K LLMs to generate the final result.
3. Experiments on benchmarks such as TruthfulQA [13], TriviaQA [8], and FACTOR [16] datasets reveal that our ensembling technique significantly enhances factual accuracy with UAF outperforming SOTA baselines by 8% in accuracy, while narrowing the performance gap with, or even surpassing, GPT-4 [1]. We also present ablation studies examining the impact of uncertainty measure, model selection criteria, and ensemble size on overall performance.

2 Proposed Method

Problem Statement: Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively, and $D = \{(x_i, y_i)\}_{i=1}^n$ the dataset, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ are the i^{th} question-answer pair. For each x_i, the goal is to generate a response $\hat{y}_i$ that maximizes the overall accuracy. The goal is to achieve this using a pool of N pretrained foundational LLMs without additional training or fine-tuning.

Algorithm 1 Components of UAF - SELECTOR and FUSER

Input: D_{val}, Pool of LLMs $\mathcal{M} = \{M^j\}_{j=1}^N$, Uncertainty function $U_f(.,.,.)$, Ensemble size K, Test data point x_{test}
Output: Test data response $\hat{y}_{test}$

procedure SELECTOR $(D_{val}, \mathcal{M}, U_f, K)$
for $M^j \in \mathcal{M}$ **do**
  for $(x_i, y_i) \in D_{val}$ **do**
    $\hat{y}_i^j = M^j(x_i) \quad ; \quad s_i^j = \mathbf{1}(\hat{y}_i^j == y_i)$
    $u_i^j = U_f(M^j, x_i, \hat{y}_i^j)$
  end for
  $Acc_j = \frac{1}{|D_{val}|}\sum_i s_i^j \; ; \quad SAH_j = ROC\_AUC\_score(\{s_i^j, u_i^j\}_i)$
  $Cscore_j = Acc_j \times SAH_j$
end for
return: $\mathcal{M}^{sel} = TopK(\{Cscore_j\}_{j=1}^N)$, {//Top K LLMs}
  $Acc^{sel} = \{Acc_j | j \in \mathcal{M}^{sel}\}$ {//Accuracy of selected K LLMs}

procedure FUSER $(x_{test}, \mathcal{M}^{sel}, Acc^{sel}, U_f, K)$
for $M^k \in \mathcal{M}^{sel}$ **do**
  $\hat{y}^k_{test} = M^k(x_{test})$
  $u^k_{test} = U_f(M^k, x_{test}, \hat{y}^k_{test})$
end for
$\hat{y}_{test} = \hat{y}^{k^*}_{test}$, where $k^* = argmax_{k \in \{1,\ldots K\}} Acc_k \times (1 - u^k_{test})$
return: $\hat{y}_{test}$

2.1 Uncertainty Aware Fusion (UAF)

Figure 2 provides an overview of our UAF framework. At a high level, UAF consists of two modules: SELECTOR and FUSER. Given a specific task, the SELECTOR selects the top K LLMs from a pool of N LLMs based on performance metric. FUSER then combines the outputs of these K LLMs to produce the final response.

2.1.1 SELECTOR. Given a pool of N LLMs denoted by $\mathcal{M}$, SELECTOR selects K LLMs (where $K < N$) to optimize computational efficiency and enhance overall factual accuracy by pruning underperforming LLMs. Selection is based on two criteria: (1) task-specific accuracy and (2) self-assessment of hallucinations based on an given uncertainty measure. Given a specified uncertainty measure $U_f(\cdot)$, and a validation set D_{val}, we prompt each LLM with input x_i, obtaining response $\hat{y}_i^j$ and corresponding uncertainty score u_i^j from the j^{th} LLM M^j. We compute the accuracy Acc_j of M^j as the fraction of correct responses. We also measure the LLMs ability for self-assessment of hallucinations SAH_j as the area under the ROC curve for the binary classification of truthful vs. hallucinatory responses using uncertainty scores. We then compute a combined score $Cscore_j = Acc_j \times SAH_j$ for each LLM. The top K models with the highest combined scores are selected greedily, where K is a hyperparameter tuned for specific tasks.

2.1.2 FUSER. Given the selected ensemble of K models $\mathcal{M}^{sel}$, for each unseen example x_{test}, we generate outputs from the K LLMs denoted by $\{\hat{y}^1_{test}, \ldots, \hat{y}^K_{test}\}$ along with the corresponding instance-specific uncertainty scores. denoted by $\{u^1_{test}, \ldots, u^K_{test}\}$. While there can be several fusion strategies, since we are dealing with natural language responses, the simplest one is to example-specific selection from the candidate outputs, i.e.,

$$\hat{y}_{test} = \hat{y}_{test}^{k^*}, \quad \text{where} \quad k^* = \arg\max_{k \in \{1,\ldots,K\}} f^k.$$

Selection criterion f^k could be based on validation set accuracy alone, inverse uncertainty or some combination of both such as $\mathrm{Acc}_k \cdot (1 - u_{test}^k)$ or $\frac{\mathrm{Acc}_k}{u_{test}^k}$. The first strategy essentially reduces the ensemble to a single most accurate model, while the second one elevates the most confident one. However, both of these approaches are sub-optimal compared to combined criteria, specifically

$$f^k = \mathrm{Acc}_k \cdot (1 - u_{test}^k),$$

which yields the best performance. Experiments with other combined selection criteria shows similar behavior to the aforementioned one and hence, we omit the results for brevity.

3 Experiments

We investigate the following questions:

- **RQ1:** How effective are uncertainty methods in detecting hallucinations?
- **RQ2:** To what extent do LLMs' relative accuracy and hallucination detection abilities vary across examples?
- **RQ3:** How does UAF perform against SOTA baselines?
- **RQ4:** How does UAF performance vary with ensemble size K?

3.1 Experimental Setup

Algorithms: We implement UAF using three uncertainty measures: Perplexity[17], Semantic Entropy[11], and Haloscope[4], each chosen to represent distinct category of approaches as mentioned in Section 1. We exclude prompting-based methods due to LLMs' tendency to be *overconfident* when verbalizing their confidence [20]. We sample 10% of the data as a validation set for our SELECTOR module and tune the hyperparameter K on this set. We compare UAF against representation editing methods: ITI [12] and TruthX [21], as well as contrastive decoding-based methods: DoLa [3] and SH2 [10]. We also evaluate UAF against three prominent ensemble-based methods: Consistent [19], MLM [18], and LLM-Blender [7].We adapt LLM-Blender for zero-shot evaluation by prompting Llama-13B to generate the final answer from the ensemble's candidate responses, instead of training the fusion model.

Datasets: For our evaluation, we consider two open-book QA datasets, TruthfulQA[13] and FACTOR-news[16], which contain 817 and 1,036 multiple-choice QA pairs, respectively, following prior works [3, 12, 21] as well as a generative QA dataset TriviaQA [8] (*rc.nocontext subset*) with 9960 QA pairs.

Metrics: For multiple-choice tasks (TruthfulQA, FACTOR), correctness is determined based on selection of the gold answer while in case of Trivia-QA, correctness is based on exact match of the ground truth. We report two metrics: `accuracy`, i.e., the fraction of correctly answered questions, and `self assessment of hallucination`, i.e., the area under ROC curve (AUROC) for detecting hallucinations using uncertainty score.

LLM-Pool: We use open-source LLMs like Llama2-13B, llama 3.2-3B, Alpaca, Vicuna, Mistralv3-7B and Opt-13B. Details of these models can be found in [5].

Table 1: AUROC of uncertainty methods in TruthfulQA. Higher values above 0.5 is preferred.

LLMs	Perplexity[17]	Haloscope[4]	Semantic Entropy[11]
Alpaca-7B	0.71	0.72	0.69
Vicuna-7B	0.8	0.87	0.82
Llama2-13B	0.68	0.71	0.72
Llama3.2-3B	0.60	0.69	0.66
Mistralv3-7B	0.51	0.78	0.63
Opt-13B	0.71	0.73	0.7

3.2 RQ1: Effectiveness of uncertainty methods in hallucination detection

We compare the performance of three uncertainty methods—Perplexity, Semantic Entropy, and Haloscope—across various LLMs in Table 1. For each example in TruthfulQA, we prompt the LLMs and mark their responses as either truthful or hallucinatory based on the ground-truth answer. We also generate a corresponding normalized uncertainty score, $\in (0, 1)$, for each response using one of the above methods. We report the area under the receiver operating characteristic curve (AUROC) in Table 1, which measures the performance of binary classification of truthful vs. hallucinatory responses by varying the thresholds of uncertainty scores. As shown in Table 1, most LLMs achieve high AUROC scores (above the random chance threshold of 0.5), indicating that uncertainty methods are effective at detecting the models' hallucinatory responses

3.3 RQ2: Variation in accuracy and hallucination detection across LLMs

We prompt five individual LLMs from our LLM pool on TruthfulQA and compute uncertainty scores using the Haloscope method for each generated response. An incorrect response is considered correctly detected as a hallucination if its uncertainty score is greater than 0.5. In Table 2, each cell in the i^{th} row and j^{th} column shows a tuple: (1) the percentage of examples where the j^{th} LLM generates the correct response missed by the i^{th} LLM, and (2) the percentage where the j^{th} LLM detects hallucinations missed by the i^{th} LLM. From this, we conclude that each LLM outperforms others in both accuracy and hallucination detection for a substantial proportion of examples. This highlights that the optimal model can vary significantly across data points, emphasizing the value of pooling strengths from multiple models in an ensemble.

Table 2: Variation in accuracy and hallucination detection capability across each LLMs in TruthfulQA.

	Alpaca-7B	Vicuna-7B	Llama2-13B	Llama3.2-3B	Mistralv3-7B
Alpaca-7B	NA	6.8 , 17	12.3 , 28.2	19 , 8.6	23.6 , 32.1
Vicuna-7B	5.6 , 21.3	NA	14.9 , 33.2	17.5 , 19.5	21.1 , 16.4
Llama2-13B	9.5 , 21	8.2 , 19.1	NA	16.6 , 10	18 , 25.4
Llama3.2-3B	8.6 , 16.6	10.2 , 24	17 , 28.7	NA	11.5 , 30
Mistralv3-7B	11 , 19.6	6.6 , 20.8	15.8 , 25.2	8.6 , 14.8	NA

Table 3: Results on TruthfulQA, TriviaQA and FACTOR-news datasets. We report accuracy (%). Best results are bolded.

	Representation Editing		Contrastive Decoding		Ensemble			UAF			GPT-4
	ITI	TruthX	DoLa	SH2	Consistent	MLM	LLM-Blender	Perplexity	Haloscope	Semantic-Entropy	(as reported [1])
TruthfulQA	38.9	54.2	33.2	34.2	41.7	29.0	48.4	56.5	**60.8**	51.3	59
TriviaQA	65.9	69.8	66.1	70.2	65.6	57.6	59	**76**	71.5	66.4	**87**
FACTOR-news	53.3	65.8	66.2	77	61.4	50.8	64.7	72.5	**81.4**	69.2	-

3.4 RQ3: UAF results on benchmark datasets

Table 3 shows UAF results across all datasets. Representation editing and contrastive decoding methods are applied to each LLM in our LLM-Pool, reporting the best result, while ensembling uses all LLMs in the pool. We implemented UAF using three different uncertainty measures: Perplexity, Haloscope, and Semantic-Entropy. UAF (with Haloscope) outperforms GPT-4 by 3% on TruthfulQA and surpasses the best baseline by 12%. In TriviaQA and Factor-news, UAF outperforms the best baseline by at least 8% and 6%, respectively, and narrows the gap with GPT-4. UAF with Semantic-Entropy performs relatively worse, likely due to the method's limited effectiveness in detecting hallucinations. This highlights that weak uncertainty modeling can hurt overall ensemble performance. Interestingly, ensembling strategies like Consistent and LLM-Blender match or outperform some complex hallucination mitigation methods, emphasizing the benefits of ensembling framework.

3.5 RQ4: Ablation study with hyperparameter K

Figure 3 shows the impact of varying K, the number of LLMs chosen by SELECTOR, on UAF accuracy across datasets. For TruthfulQA, the highest accuracy (60.8%) is achieved with top 5 LLMs (Mistralv3-7B, Vicuna, Llama3.2-3B, Opt-13B, and Llama2-13B). For TriviaQA and FACTOR, the top 4 LLMs (Llama2-13B, Mistralv3-7B, Alpaca, Llama3.2-3B) and top 3 LLMs (Mistralv3-7B, Llama3.2-3B, Llama2-13B) yield optimal results, with performance sharply declining for larget K. These results highlight the importance of the SELECTOR module in the UAF method, demonstrating that carefully choosing a small subset of high-performing LLMs enhances ensemble performance more effectively than using the entire LLM pool.

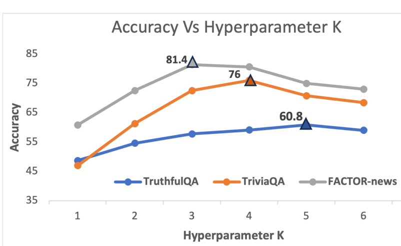

Figure 3: Impact of K on UAF performance.

4 Conclusion

In this work, we introduced an ensemble framework UAF that reduces hallucinations in factoid question answering (QA) by leveraging both the accuracy and self-assessment capabilities of multiple LLMs. By incorporating uncertainty estimation, UAF strategically combines model responses, improving factual accuracy on multiple benchmark datasets. Since UAF combines the responses of multiple LLMs, the inference time scales linearly with the number of models in the ensemble, requiring multiple forward passes through each model to generate responses and uncertainty estimates. Additionally, the uncertainty estimation itself may introduce some overhead, depending on the specific method used to compute it. Despite the added computational complexity of ensemble models, the significant improvement in factual accuracy justifies the approach. Future work could explore dynamic LLM selection based on each data point and integrate reinforcement learning for better adaptation to diverse data, enhancing both accuracy and efficiency.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[2] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. *arXiv preprint arXiv:2402.03744* (2024).
[3] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).
[4] Xuefeng Du and Chaowei Xiao. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *arXiv preprint arXiv:2409.17504* (2024).
[5] Guangyu Hou and Qin Lian. 2024. Benchmarking of Commercial Large Language Models: ChatGPT, Mistral, and Llama. (2024).
[6] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
[7] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561* (2023).
[8] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).
[9] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).
[10] Jushi Kai, Tianhang Zhang, and Zhouhan Lin. 2024. SH2: Self-Highlighted Hesitation Helps You Decode Truthfully. *arXiv preprint arXiv:2401.05930* (2024).
[11] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664* (2023).
[12] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems* 36 (2024).
[13] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).
[14] Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650* (2020).
[15] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).
[16] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908* (2023).
[17] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective

generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

[18] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659* (2019).

[19] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[20] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063* (2023).

[21] Shaolei Zhang, Tian Yu, and Yang Feng. 2024. TruthX: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811* (2024).