Decoding Uncertainty: The Impact of Decoding Strategies for Uncertainty Estimation in Large Language Models

Wataru Hashimoto, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology
{hashimoto.wataru.hq3, kamigaito.h, taro}@is.naist.jp

Abstract

Decoding strategies manipulate the probability distribution underlying the output of a language model and can therefore affect both generation quality and its uncertainty. In this study, we investigate the impact of decoding strategies on uncertainty estimation in Large Language Models (LLMs). Our experiments show that Contrastive Search, which mitigates repetition, yields better uncertainty estimates on average across a range of preference-aligned LLMs. In contrast, the benefits of these strategies sometimes diverge when the model is only post-trained with supervised fine-tuning, i.e. without explicit alignment.

1 Introduction

Recent advances in natural language processing (NLP) have been driven almost entirely by the rapid progress of Large Language Models (LLMs). State-of-the-art models such as GPT-4 (OpenAI et al., 2023), Llama (Touvron et al., 2023a), and DeepSeek (DeepSeek-AI et al., 2025) already match or surpass human performance on a diverse suite of downstream NLP tasks.

Despite these successes, LLMs sometimes output fabricated or misleading text (hallucinations), which hinders the deployment of LLMs in safety-critical domains such as medicine, finance, and law. Uncertainty Estimation (UE) is a key technique for mitigating the problem (Geifman et al., 2019; Galil et al., 2023; Xin et al., 2021; Hashimoto et al., 2024). By quantifying predictive uncertainty, a system can reject dubious outputs and route them to either human experts or stronger models.

In addition, decoding strategies also constitute a promising approach to address the problem. Decoding strategies manipulate next-token distributions of language models, thereby able to elicit higher-quality outputs from the language model. Recent work demonstrates that the choice of decoding strategy can markedly impact the quality of LLM outputs, underscoring its pivotal role in unlocking the full potential of these models (Shi et al., 2024).

However, the comprehensive investigation into how decoding strategies affect UE performance in LLMs remains limited. Although recent studies improve uncertainty by devising sampling strategies (Aichberger et al., 2025; Vashurin et al., 2025), the UE performance combined with extensive decoding strategies has not been systematically evaluated across various tasks. Since decoding algorithms influence both the probability distribution over candidate tokens and the final token selection, they can have a significant impact on UE performance. Furthermore, mainstream LLMs usually apply *preference-alignment* techniques including Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) or Direct Preference Optimization (DPO) (Rafailov et al., 2023) after Supervised Fine-Tuning (SFT). Although such techniques improve the alignment of outputs with human preferences, recent work suggests they can degrade reliability (Kadavath et al., 2022; OpenAI et al., 2023; Tian et al., 2023; Xiao et al., 2025), potentially interacting with the choice of decoding strategy. The investigation of these interactions is therefore essential for producing more trustworthy LLM outputs. We address this gap through two research questions:

- RQ1: Which decoding strategies deliver the best UE performance?
- RQ2: How do training stages such as SFT and the preference-alignment techniques modulate UE performance across decoding strategies?

Our experiments reveal the following findings: First, Contrastive Search, which explicitly mitigates repetition, achieves better UE performance as a whole. Second, the optimal decoding strategy for UE can change as a model pro-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14601–14613
November 4-9, 2025

gresses from SFT to preference alignment during its post-training phase. In addition, our results show that the changes in UE performance depend on the interaction with the decoding strategy and preference-alignment techniques. All code is available at https://github.com/wataruhashimoto52/decoding_uncertainty.

2 Decoding Strategies

We focus exclusively on deterministic decoding strategies because deterministic outputs are important in safety-critical domains such as finance (Bender et al., 2021; You and Chon, 2024). The strategies examined in this study are: Greedy Search (Greedy), Beam Search (BS) (Freitag and Al-Onaizan, 2017), Diverse Beam Search (DBS) (Vijayakumar et al., 2018), Contrastive Search (CS) (Su et al., 2022; Su and Collier, 2023), Contrastive Decoding (CD) (Li et al., 2023), Frustratingly Simple Decoding (FSD; based on an n-gram model) (Yang et al., 2024), FSD-vec (based on a vectorized n-gram model), Decoding by Contrastive Layers (DoLa) (Chuang et al., 2024), and Self-Logits Evolution Decoding (SLED) (Zhang et al., 2024a). Technical details and the hyper-parameter search space are provided in Appendix A.

3 Experimental Settings

3.1 Datasets

We conducted evaluations across four text generation tasks: question answering (QA), text summarization (TS), machine translation (MT), and code generation (CG). In QA, we use TriviaQA (Joshi et al., 2017) dataset. In TS, we use XSum (Narayan et al., 2018) dataset. In MT, we use WMT19 (Foundation, 2019) dataset in German to English (De-En) setting. In CG, we use HumanEval (Chen et al., 2021) dataset. Dataset details are in Appendix B.

3.2 Models

In RQ1, to examine the impact of decoding methods on UE performance across multiple tasks, we used Llama2-7B-Chat (Touvron et al., 2023b)[1], Llama3-8B-RLHF (Grattafiori et al., 2024; Hu et al., 2024)[2], and Zephyr-7B-β (Tunstall et al., 2024).[3] For CD, we adopted TinyLlama (Zhang et al., 2024b)[4] as the amateur model. In RQ2, to evaluate the effects of the SFT and RLHF stages, we employed Llama3-8B-SFT[5] and Llama3-8B-RLHF. When applying preference tuning, we used Llama3-8B-DPO[6], which is applied the iterative version of DPO (Dong et al., 2024). For CD, we adopted Llama3.2-1B-Instruct[7] as the amateur.

3.3 Details of Uncertainty Estimation

3.3.1 Uncertainty Estimation Metrics

Following Fadeeva et al. (2023), we measure UE performance with the Prediction–Rejection Ratio (PRR), which compares the area under the prediction–rejection curve obtained when ranking generations by model uncertainty to the oracle curve that ranks by true quality. Unlike AUROC, PRR does not require binary labels, making it applicable to various text generation tasks such as TS or MT. Let the test set be $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$. For each input $\boldsymbol{x}_i$, the language model produces an output $f(\boldsymbol{x}_i)$ and an associated uncertainty score $\mathcal{U}(\boldsymbol{x}_i)$. The Prediction–Rejection Curve (PRC) traces the average min-max normalized generation quality $\mathcal{Q}(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$ of those outputs that satisfy $\mathcal{U}(\boldsymbol{x}_i) < a$ as the rejection threshold a varies. The PRR compares the area under this curve when ranking by uncertainty against an oracle that ranks by true quality:

$$PRR = \frac{PRC_{\text{uns}}}{PRC_{\text{orc}}}. \quad (1)$$

Here, PRC_{orc} is the area obtained when the lowest-quality samples are rejected first, whereas PRC_{uns} is the area when rejection is driven by the model's uncertainty scores. Because uncertainty is an imperfect proxy for quality, PRC_{uns} typically lies below PRC_{orc}. A higher PRR means the uncertainty scores more accurately filter out low-quality outputs.[8]

The quality score $\mathcal{Q}$ is task-dependent. The quality scores used in calculating the PRR are as follows: for QA we use RougeL (Lin, 2004); for

[1] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[2] https://huggingface.co/OpenRLHF/Llama-3-8b-rlhf-100k

[3] https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

[4] https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-955k-token-2T

[5] https://huggingface.co/OpenRLHF/Llama-3-8b-sft-mixture

[6] https://huggingface.co/RLHFlow/LLaMA3-iterative-DPO-final

[7] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[8] If correctly predicted instances receive higher uncertainty than mispredicted ones, the PRR can become negative.

Model	Method	MSP							MTE						
		TriviaQA	XSum		WMT19			HumanEval	TriviaQA	XSum		WMT19			HumanEval	
		RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1	Mean PRR	RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1	Mean PRR
Llama2-7B-Chat	Greedy	$62.97_{0.59}$	$14.42_{1.87}$	$1.57_{1.78}$	$38.74_{2.40}$	$46.48_{1.92}$	$19.02_{3.14}$	$-11.03_{8.61}$	20.95	$49.13_{0.91}$	$8.07_{1.60}$	$10.68_{1.75}$	$31.24_{2.21}$	$25.03_{1.93}$	$21.69_{3.33}$	$-13.49_{8.56}$	17.89
	BS	$63.62_{0.70}$	$14.12_{1.79}$	$-0.45_{1.66}$	$38.96_{2.02}$	$52.88_{1.56}$	$20.16_{3.56}$	$-24.51_{7.04}$	18.65	$50.58_{0.85}$	$4.68_{1.75}$	$9.50_{1.69}$	$29.78_{2.10}$	$21.46_{2.05}$	$18.11_{3.71}$	$-28.47_{6.56}$	14.03
	DBS	$63.98_{2.34}$	$14.44_{1.71}$	$-18.03_{1.42}$	$28.97_{2.41}$	$53.24_{1.82}$	$4.35_{3.66}$	$-35.36_{8.70}$	21.51	$41.34_{0.94}$	$7.50_{1.64}$	$12.54_{1.76}$	$6.66_{2.34}$	$-6.66_{2.27}$	$6.07_{3.44}$	$-11.39_{8.38}$	14.25
	CS	$63.73_{0.69}$	$17.56^{*}_{1.04}$	$1.43_{2.04}$	$36.94_{2.29}$	$41.99_{1.97}$	$19.58_{3.21}$	$-9.96_{8.86}$	21.55	$51.95^{*}_{0.82}$	$8.99_{1.85}$	$10.21_{2.14}$	$30.66_{2.11}$	$23.70_{2.09}$	$22.24_{3.13}$	$-12.10_{8.28}$	18.66
	CD	$15.97_{2.22}$	$7.18_{1.57}$	$-1.22_{1.19}$	$49.31^{*}_{1.43}$	$54.19^{*}_{1.34}$	$50.39^{*}_{2.14}$	$-7.20_{8.44}$	21.47	$24.95_{1.67}$	$9.33_{1.68}$	$1.47_{2.04}$	$58.25^{*}_{1.27}$	$62.48^{*}_{1.26}$	$63.00^{*}_{1.87}$	$-5.67_{8.08}$	27.11
	FSD	$33.84_{1.87}$	$0.64_{1.39}$	$12.42^{*}_{2.27}$	$31.82_{1.74}$	$14.04_{2.19}$	$8.15_{3.58}$	$-27.03_{8.42}$	9.97	$34.81_{0.95}$	$-0.13_{1.38}$	$11.73_{2.26}$	$34.77_{1.77}$	$15.93_{2.17}$	$9.39_{3.55}$	$-27.03_{8.29}$	10.59
	FSD-vec	$32.66_{1.81}$	$-1.58_{1.36}$	$10.11_{2.14}$	$31.24_{1.71}$	$12.83_{1.71}$	$8.24_{3.45}$	$-20.19_{8.71}$	16.81	$32.08_{0.72}$	$-3.04_{1.36}$	$11.31_{2.01}$	$33.54_{1.78}$	$14.21_{2.15}$	$9.10_{3.44}$	$-9.14_{8.51}$	15.36
	DoLa	$61.15_{0.65}$	$12.46_{1.92}$	$0.07_{1.56}$	$38.78_{2.04}$	$49.15_{1.79}$	$16.74_{3.62}$	$-14.82_{8.70}$	19.06	$49.23_{0.85}$	$5.51_{1.96}$	$8.17_{1.72}$	$25.03_{1.89}$	$17.36_{1.91}$	$15.14_{3.67}$	$-17.41_{9.02}$	14.28
	SLED	$-12.68_{0.44}$	-	-	$-27.36_{1.20}$	$-61.37_{1.54}$	$2.53_{3.04}$	$-41.24_{8.58}$	-19.69	$7.80_{0.38}$	-	-	$16.24_{1.96}$	$18.32_{2.06}$	$5.34_{3.24}$	$-17.63_{9.78}$	2.94
Llama3-8B-RLHF	Greedy	$20.80_{0.86}$	$15.23_{1.89}$	$-2.44_{1.77}$	$59.6_{1.54}$	$82.27_{0.72}$	$20.49_{2.89}$	$-30.76_{8.18}$	23.60	$20.27_{0.87}$	$9.78_{1.62}$	$5.79_{1.93}$	$59.65_{1.37}$	$68.40_{1.01}$	$23.96_{2.75}$	$-31.33_{8.26}$	22.36
	BS	$22.88^{*}_{0.80}$	$38.32^{*}_{1.33}$	$2.64_{1.61}$	$31.36_{1.46}$	$48.11_{1.64}$	$-6.71_{4.23}$	$-38.85_{5.46}$	13.96	$19.31_{0.85}$	$-1.54_{1.72}$	$2.27_{1.85}$	$25.06_{2.20}$	$8.10_{2.30}$	$16.87_{3.46}$	$-41.03_{1.87}$	4.15
	DBS	$-9.70_{0.71}$	$12.03_{1.93}$	$-9.52_{1.66}$	$29.8_{1.44}$	$24.84_{1.22}$	$16.20_{3.92}$	$26.01^{*}_{10.17}$	12.81	$11.68_{0.81}$	$8.95_{1.82}$	$8.42_{2.09}$	$13.48_{1.85}$	$15.77_{2.13}$	$5.66_{4.65}$	$-39.76_{5.40}$	3.46
	CS	$16.89_{0.84}$	$16.37_{1.83}$	$-2.51_{1.53}$	$61.24^{*}_{1.53}$	$84.86^{*}_{0.66}$	$19.56_{3.10}$	$-28.34_{9.62}$	24.01	$19.20_{0.79}$	$11.96^{*}_{1.57}$	$4.44_{1.73}$	$62.93^{*}_{1.38}$	$72.79^{*}_{0.98}$	$22.54_{2.97}$	$-29.79_{8.63}$	23.44
	CD	$17.81_{0.96}$	$-9.18_{1.54}$	$9.90^{*}_{2.15}$	$-9.03_{1.96}$	$-57.07_{1.42}$	$-34.46_{2.67}$	$-51.43_{5.84}$	-19.07	$22.25^{*}_{0.80}$	$-0.05_{1.84}$	$9.33^{*}_{1.83}$	$41.62_{1.54}$	$25.74_{1.67}$	$60.24^{*}_{2.01}$	$-51.70_{5.85}$	15.35
	DoLa	$20.11_{0.87}$	$12.27_{2.02}$	$-3.29_{1.51}$	$50.31_{1.82}$	$76.96_{1.22}$	$21.13_{3.02}$	$-39.79_{5.85}$	19.67	$18.41_{0.80}$	$6.29_{1.78}$	$4.63_{1.77}$	$52.17_{1.80}$	$64.86_{1.11}$	$19.90_{2.75}$	$-40.46_{5.96}$	17.97
Zephyr-7B-β	Greedy	$64.85_{0.55}$	$11.80_{2.05}$	$-5.63_{1.79}$	$65.53_{1.53}$	$81.67_{0.50}$	$22.36_{3.31}$	$-27.41_{5.56}$	30.45	$53.79_{0.77}$	$10.80_{1.85}$	$3.19_{2.09}$	$64.82_{1.46}$	$72.41_{0.89}$	$17.98^{*}_{3.63}$	$-35.24_{5.01}$	26.82
	BS	$53.29_{1.18}$	$10.45_{1.92}$	$-5.32_{1.21}$	$59.87_{1.71}$	$80.85_{0.81}$	$22.01_{3.41}$	$-35.43_{4.71}$	26.53	$44.48_{1.16}$	$10.06_{1.73}$	$2.13_{2.05}$	$57.41_{1.57}$	$62.98_{1.17}$	$12.02_{3.96}$	$-31.31_{3.67}$	22.54
	DBS	$64.19_{0.61}$	$1.50_{1.82}$	$-17.84_{2.39}$	$43.86_{1.45}$	$75.24_{0.93}$	$21.52_{3.16}$	$-21.22^{*}_{7.67}$	23.89	$24.42_{0.88}$	$10.16_{1.57}$	$10.74^{*}_{1.94}$	$36.38_{1.96}$	$50.67_{1.49}$	$16.28_{3.66}$	$-17.96^{*}_{5.92}$	18.67
	CS	$65.22^{*}_{0.51}$	$12.20^{*}_{1.97}$	$-5.34_{1.83}$	$65.77_{1.47}$	$86.02^{*}_{0.58}$	$23.72^{*}_{3.30}$	$-27.60_{7.18}$	31.43	$54.29^{*}_{0.77}$	$11.67_{1.84}$	$4.22_{2.16}$	$65.00^{*}_{1.45}$	$72.94^{*}_{1.06}$	$17.48_{3.49}$	$-28.63_{6.95}$	28.14
	FSD	$23.39_{0.83}$	$11.27_{2.26}$	$2.04^{*}_{2.48}$	$42.28_{1.78}$	$24.44_{1.76}$	$11.52_{3.13}$	$-34.39_{5.28}$	11.51	$23.87_{0.86}$	$11.62_{2.23}$	$3.32_{2.56}$	$44.97_{1.76}$	$27.66_{1.76}$	$12.54_{3.05}$	$-36.03_{1.88}$	12.56
	FSD-vec	$23.21_{0.83}$	$11.82_{2.12}$	$-6.57_{2.54}$	$41.23_{1.66}$	$26.66_{1.79}$	$15.39_{3.12}$	$-28.12_{6.51}$	15.03	$23.85_{0.86}$	$14.35^{*}_{2.15}$	$2.79_{2.52}$	$43.84_{1.66}$	$28.64_{1.73}$	$16.55_{2.98}$	$-30.40_{6.49}$	14.23
	DoLa	$63.90_{0.57}$	$11.40_{2.04}$	$-5.36_{1.78}$	$65.22_{1.55}$	$84.54_{0.54}$	$22.81_{3.24}$	$-26.99_{5.69}$	30.79	$52.02_{0.80}$	$10.35_{1.84}$	$2.30_{2.08}$	$60.00_{1.62}$	$65.76_{1.12}$	$15.82_{3.25}$	$-28.72_{5.29}$	25.36

Table 1: PRRs for every task and generation metric pair in Llama2-7B-Chat, Llama3-8B-RLHF, and Zephyr-7B-β. Warmer color indicates better results. * indicates that the best strategy is significantly better ($p < 0.05$) than the second best. All standard deviations are obtained by bootstrap resampling with 1,000 trials.

Model	Method	TriviaQA	XSum		WMT19			HumanEval
		RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Llama2-7B-Chat	Greedy	8.43	10.57	10.57	6.14	6.14	6.14	8.52
	BS	24.95	7.41	7.13	16.26	16.26	20.30	6.26
	DBS	1,751.65	2,020.93	2,012.05	1,971.69	1,971.69	1,977.82	2,604.76
	CS	17.42	10.61	10.61	8.16	8.16	6.11	12.77
	CD	62.97	186.72	186.72	63.48	63.74	63.36	98.20
	FSD	153.70	15.52	16.25	8.96	8.96	8.96	7.56
	FSD-vec	92.10	15.57	16.99	9.09	9.09	8.92	102.44
	DoLa	6.32	7.54	7.33	4.82	4.82	4.82	6.69
	SLED	-1,720.95	-1,959.09	-2,441.84	-2,669.11	-2,135.69	-2,669.11	-4,218.55
Llama3-8B-RLHF	Greedy	34.76	9.00	9.00	9.00	23.35	23.35	17.13
	BS	90.52	351.83	351.83	351.83	187.62	191.84	8.24
	DBS	2249.43	1210.59	1175.13	1213.68	1537.93	1609.42	2488.23
	CS	50.10	9.58	9.58	9.58	23.42	24.19	26.49
	CD	47.81	175.62	175.62	175.62	374.31	374.31	10.68
	DoLa	24.43	4.91	4.89	4.89	15.33	15.33	10.63
Zephyr-7B-β	Greedy	15.79	20.70	20.70	11.12	11.12	11.12	8.11
	BS	41.49	17.81	16.91	19.47	19.47	19.47	5.17
	DBS	1,470.62	1,958.16	1,958.16	1,276.36	1,276.36	1,395.72	2,210.45
	CS	16.03	21.45	21.45	11.46	19.08	11.46	9.58
	FSD	23.29	23.30	23.30	18.69	20.23	20.23	9.73
	FSD-vec	23.28	23.62	23.62	18.62	20.25	20.25	11.72
	DoLa	12.63	16.73	16.73	9.01	9.01	9.30	6.67

Table 2: Averaged MSP scores for every task and generation metric pair in Llama2-7B-Chat, Llama3-8B-RLHF, and Zephyr-7B-β. Higher score indicates more uncertain.

Model	Method	TriviaQA	XSum		WMT19			HumanEval
		RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Llama2-7B-Chat	Greedy	0.13	0.16	0.16	0.25	0.25	0.25	0.10
	BS	0.10	0.32	0.26	0.23	0.23	0.23	0.08
	DBS	0.21	0.37	0.35	0.16	0.16	0.25	0.19
	CS	0.13	0.40	0.40	0.25	0.25	0.25	0.12
	CD	0.25	0.54	0.54	0.16	0.27	6.03	0.09
	FSD	0.11	0.60	0.60	0.25	0.25	0.25	0.09
	FSD-vec	0.24	0.55	0.55	0.25	0.25	0.25	0.09
	DoLa	0.05	0.19	0.23	0.18	0.18	0.18	0.05
Llama3-8B-RLHF	Greedy	0.80	0.36	0.36	0.36	0.89	0.89	0.39
	BS	0.52	0.27	0.27	0.27	0.54	0.51	0.18
	DBS	0.48	0.46	0.39	0.34	0.59	0.60	0.44
	CS	0.75	0.37	0.37	0.37	0.88	0.88	0.45
	CD	0.85	0.74	0.74	0.74	0.76	0.76	0.24
	DoLa	0.41	0.19	0.19	0.19	0.42	0.42	0.18
Zephyr-7B-β	Greedy	0.44	0.50	0.50	0.34	0.34	0.34	0.18
	BS	0.30	0.41	0.38	0.25	0.25	0.25	0.11
	DBS	0.38	0.45	0.45	0.29	0.29	0.33	0.27
	CS	0.44	0.49	0.49	0.34	0.34	0.34	0.20
	FSD	0.52	0.51	0.51	0.42	0.44	0.44	0.21
	FSD-vec	0.52	0.51	0.51	0.42	0.43	0.43	0.22
	DoLa	0.26	0.31	0.31	0.21	0.21	0.21	0.12

Table 3: Averaged MTE scores for every task and generation metric pair in Llama2-7B-Chat, Llama3-8B-RLHF, and Zephyr-7B-β. Higher score indicates more uncertain.

TS we report RougeL and AlignScore (Zha et al., 2023); for MT we report BLEU (Papineni et al., 2002), Comet (Rei et al., 2020) and AlignScore; for CG we report Pass@1 (Chen et al., 2021). To improve readability, all generation quality scores and PRRs are multiplied by 100.

3.3.2 How to Estimate Uncertainty Score

To convert the predictive token-level probability distribution into a single uncertainty score, an aggregation scheme must be chosen. To analyze the impact of decoding strategies on predictive uncertainty from the viewpoint of probability and entropy, we limit our analysis to two fundamental methods: Maximum Sequence Probability (MSP) which is the negative log-likelihood of the generated sequence, and Mean Token Entropy (MTE) which is the average entropy of the token-level predictive distributions.[9] For each decoding strategy tested, MSP and MTE are computed, and the resulting uncertainty scores are then evaluated using PRRs.

4 Results & Analysis

4.1 RQ1: Which decoding strategies deliver the best UE performance?

Table 1 reports the PRRs obtained with MSP and MTE when each of the decoding strategies is applied across four benchmarks. In addition, Table 2 and Table 3 report averaged MSP scores and MTE scores, respectively.

[9] More advanced methods can affect uncertainty estimates combined with the choice of the decoding strategy. Results obtained by applying one such technique – Shifting Attention to Relevance (Duan et al., 2024) – to the computation of uncertainty scores are presented in Appendix H.

Model	Method	Distinct-1				Distinct-2			
		TriviaQA	XSum	WMT19	HumanEval	TriviaQA	XSum	WMT19	HumanEval
Llama2-7B-Chat	Greedy	0.750	0.809	0.784	0.608	0.924	0.974	0.935	0.851
	BS	0.744	0.808	0.775	0.602	0.919	0.973	0.929	0.845
	DBS	0.739	0.806	0.766	0.598	0.917	0.973	0.925	0.842
	CS	0.759	0.819	0.807	0.604	0.930	0.977	0.948	0.846
	CD	0.739	0.782	0.729	0.569	0.909	0.950	0.887	0.801
	FSD	0.735	0.773	0.716	0.573	0.910	0.946	0.884	0.808
	FSD-vec	0.734	0.766	0.706	0.556	0.907	0.942	0.881	0.785
	DoLa	0.748	0.771	0.716	0.560	0.923	0.946	0.888	0.792
	SLED	0.744	-	0.716	0.554	0.919	-	0.889	0.790
Llama3-8B-RLHF	Greedy	0.703	0.805	0.766	0.656	0.914	0.941	0.861	0.885
	BS	0.692	0.804	0.766	0.648	0.909	0.941	0.861	0.881
	DBS	0.677	0.783	0.735	0.642	0.899	0.924	0.828	0.877
	CS	0.740	0.866	0.864	0.666	0.943	0.991	0.965	0.890
	CD	0.687	0.750	0.689	0.650	0.905	0.883	0.789	0.882
	DoLa	0.678	0.766	0.678	0.644	0.907	0.901	0.799	0.881
Zephyr-7B-β	Greedy	0.734	0.778	0.755	0.545	0.920	0.963	0.887	0.792
	BS	0.754	0.776	0.782	0.531	0.919	0.960	0.896	0.772
	DBS	0.747	0.775	0.771	0.538	0.919	0.960	0.890	0.787
	CS	0.734	0.782	0.774	0.557	0.922	0.966	0.906	0.808
	FSD	0.666	0.762	0.648	0.544	0.895	0.957	0.827	0.793
	FSD-vec	0.665	0.760	0.644	0.546	0.892	0.956	0.816	0.797
	DoLa	0.734	0.778	0.753	0.545	0.920	0.963	0.886	0.792

Table 4: Distinct-1 and Distinct-2 for every task and generation metric pair in Llama2-7B-Chat, Llama3-8B-RLHF, and Zephyr-7B-β. Higher score indicates diversified outputs.

Contrastive Search shows better uncertainty across the models on average. Across all aligned models examined, CS, followed by Greedy, produces better uncertainty, on average. We hypothesized that these results are due to CS's ability to mitigate repetition which is one of the causes of overconfidence in a language model (Holtzman et al., 2020) while keeping the original probability (Su et al., 2022; Su and Collier, 2023). To evaluate this, we measured averaged sentence-level Distinct-n (Li et al., 2016). Distinct-n is the rate at which n-grams in the output are different, which can evaluate the diversity of tokens in the output sentences. As shown in Table 4, CS has the highest Distinct-1 and Distinct-2 overall, suggesting that the outputs from CS have less repetition than other decoding strategies.

BS and DBS sometimes underperform. We can see that BS and DBS in Llama3-8B-RLHF and Zephyr-7B-β sometimes perform worse. In Tables 2 and Table 3, the negative log-probability and the entropy change significantly with BS and DBS compared to Greedy or CS. The uncertainty scores obtained based on the manipulated probability distributions by BS and DBS may not be aligned with the objective of separating high- and low-quality outputs.

For CD, there is a large difference in UE performance across models. On average, CD provides the strongest aggregate performance, followed by CS in Llama2-7B-Chat. However, the advantage of CD is mainly pronounced in MT setting. On the other hand, in Llama3-8B-RLHF, the situation is markedly different: while CD remains reasonably reliable for factual metrics, its reliability deteriorates sharply for MT or TS. As CD is highly sensitive to the specific pairing of teacher and student models, substantial behavioral differences across model families can be expected. Results of other models also support this in Appendix E and Appendix F. Moreover, as with BS and DBS, we can see that the probability distribution of the LLM output changes significantly when CD is used, from Table 2 and Table 3. The selection and construction of an appropriate amateur model for CD to optimize UE performance remains an open challenge.

Recent factuality decoding strategies underperform. As shown in Table 1, recent factuality decoding strategies such as DoLa and SLED frequently underperform alternative methods in terms of PRRs. Factuality decoding is hypothesized to increase factual correctness by amplifying knowledge that is localised within particular layers of the language model. This amplification, however, can distort the probability distribution of the base LLMs, potentially degrading downstream performance. The results in Table 2 and Table 3 reveal that factuality decoding strategies provide overconfident MSP score and less entropy, suggesting that the original probability distribution of the language model is indeed being altered by emphasising factual tokens.

4.2 RQ2: How do training stages such as SFT and the preference-alignment techniques modulate UE performance across decoding strategies?

Stopping post-train at the SFT stage may affect the conclusion of RQ1. Figure 1 depicts the change of PRR values across all task-quality pairs when the training phase of Llama3-8B is switched from SFT to RLHF. The figure reveals that under SFT, BS achieves superior PRR in a larger number of cases than it does under RLHF. Previous research (Kumar and Sarawagi, 2019) has shown that the confidence calibration effect of BS has positive impacts not only the confidence but also the generation quality. Furthermore, applying RLHF to an SFT model tends to make its token-level probability distribution more overconfident (Xie et al., 2024). Consequently, during beam search, low uncertainty score can be assigned to low quality outputs. This miscalibration will lead to a degradation in PRR.

The absolute impact of the training stage transition on PRR is task-dependent. On TriviaQA, applying RLHF reduces PRR, whereas on WMT19

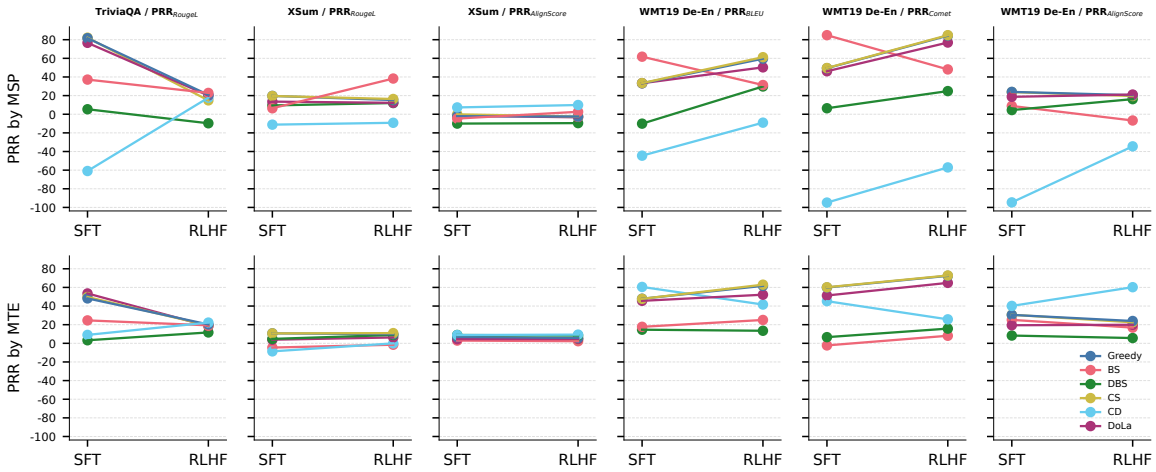

Figure 1: Slopegraphs of PRR when changing the model from Llama3-8B-SFT to Llama3-8B-RLHF.

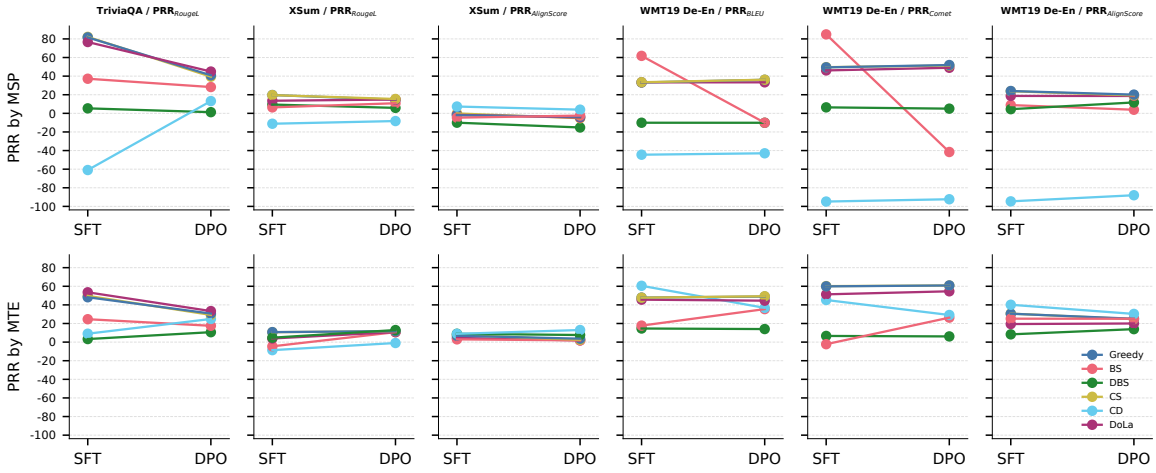

Figure 2: Slopegraphs of PRR when changing the model from Llama3-8B-SFT to Llama3-8B-DPO.

De-En it tends to enhance PRR. Overall, we did not observe the tendency that RLHF induces overconfidence and reduces predictive uncertainty (Kadavath et al., 2022; Xie et al., 2024) when PRR was used as the evaluation metric.

On the other hand, as shown in Figure 2, there are fewer cases in which PRR improved when applied DPO compared to Figure 1, suggesting overconfident than Llama3-8B-RLHF. The reason why DPO has lower UE performance compared to RLHF can be due to an interesting "squeezing effect" (Ren and Sutherland, 2025) in the training of Llama3-8B-DPO. The "squeezing effect" happens a concentration of probability mass on the most likely token by the negative gradient when using DPO-like loss, while Proximal Policy Optimization (PPO) (Schulman et al., 2017) loss in RLHF avoids the effect (Ren and Sutherland, 2025). The phenomenon that Greedy or CS improved PRR in MT on the RLHF stage was not observed on the DPO stage, which can also be due to the increase in the probability of the most likely token and the degeneration of the probability other than the most likely token.

5 Conclusion

In this study, we examined how decoding strategies affect predictive uncertainty in LLMs. Our experiments show that Contrastive Search strategy tends to provide better uncertainty estimates across various tasks and models on average by mitigating output repetition, a key source of model overconfidence. On the other hand, we found that the conclusions may change depending on the stage in the post-training phase, such as SFT and the preference-alignment. We hope that this study will help practitioners improve the reliability of LLMs.

Limitations

LLMs Our study mainly relies on two aligned models (Llama2-7B-Chat and Llama3-8B-RLHF) and a single SFT model (Llama3-8B-SFT). Experiments on larger models are limited by available resources. In addition, proprietary models such as GPT4 (OpenAI et al., 2023) or Gemini series (Team et al., 2024, 2025) are black boxes. Therefore, users cannot freely manipulate the decoding strategy. All experiments fix the prompt template; we do not explore how prompt engineering might change the conclusions.

Decoding Strategies From the viewpoint of practice, we did not consider stochastic decoding strategies, which cannot guarantee deterministic outputs, as discussed in Section 2. In addition, our small-scale experiments do not suggest that stochastic decoding strategies are effective for PRR (see Appendix G). However, we may find the better stochastic strategy in terms of PRR by more extensive experiments. Moreover, some recent decoding strategies such as ϕ-Decoding (Xu et al., 2025) are omitted for limited resource reasons.

UE Methods Our analysis focuses on two classical, token-probability based uncertainty estimators - MSP and MTE. More advanced techniques such as Semantic Entropy (Kuhn et al., 2023), Shifting Attention to Relevance (Duan et al., 2024), and distance-based methods (Yoo et al., 2022; Hashimoto et al., 2025) are not benchmarked systematically. As a result, we cannot claim that the decoding strategy ranking we report would persist when paired with stronger uncertainty estimators.

Tasks Our benchmark suite covers four English-only generation tasks with public test sets. Tasks such as multi-modal understanding (Yue et al., 2024), combining Retrieval Augmented Generation (RAG) setting (Ozaki et al., 2025), and non-English setting (Raihan et al., 2025) are out of scope. By improving comprehensiveness, we are likely to gain a deeper understanding of the strengths and weaknesses of each decoding strategy.

Ethical Considerations

AI Assistant Tools We used ChatGPT[10] and GitHub Copilot[11] to accelerate our research.

Datasets & Models This study relies exclusively on publicly available datasets (TriviaQA, XSum, WMT19 De→En, and HumanEval) and openly released LLMs. All datasets and LLMs used in this study are, at a minimum, licensed for research purposes. In addition, the datasets we used do not consist of harmful domains (see Appendix B).

Uncertainty Estimation Even a high PRR score can miss low-quality generations; therefore, critical decisions must always include qualified human oversight.

Acknowledgements

The authors also acknowledge the Nara Institute of Science and Technology's HPC resources made available for conducting the research reported in this paper.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

[10] https://openai.com/index/chatgpt/

[11] https://github.com/features/copilot

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Wikimedia Foundation. 2019. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. 2023. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *The Eleventh International Conference on Learning Representations*.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*, pages 18852–18867, Miami, Florida, USA. Association for Computational Linguistics.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. 2025. Efficient nearest neighbor based uncertainty estimation for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4350–4366, Albuquerque, New Mexico. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2025. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 1–17, Viena, Austria. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Md Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mHumanEval - a multilingual benchmark to evaluate large language models for code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11432–11461, Albuquerque, New Mexico. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yi Ren and Danica J. Sutherland. 2025. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation. *Transactions on Machine Learning Research*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Cocoa: A generalized approach to uncertainty quantification by integrating confidence and consistency of llm outputs. *arXiv preprint arXiv:2502.04964*.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Forty-second International Conference on Machine Learning*.

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*, pages 18128–18138, Miami, Florida, USA. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. 2025. ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13214–13227, Vienna, Austria. Association for Computational Linguistics.

Haoran Yang, Deng Cai, Huayang Li, Wei Bi, Wai Lam, and Shuming Shi. 2024. A frustratingly simple decoding method for neural text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 536–557, Torino, Italia. ELRA and ICCL.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Doohee You and Dan Chon. 2024. Trust & safety of llms and llms in trust & safety. *arXiv preprint arXiv:2412.02113*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024a. SLED: Self logits evolution decoding for improving factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024b. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

A Details of Decoding Strategies

Greedy Search (Greedy) is the simplest decoding strategy, where at each time step t the token with the highest conditional probability is selected. Formally, given an input $\boldsymbol{x}$ and the previously generated sequence $\boldsymbol{y}_{<t} = \{y_1, y_2, \ldots, y_{t-1}\}$, the next token y_t is chosen as:

$$y_t = \underset{y \in \mathcal{V}}{\operatorname{argmax}} P(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \tag{2}$$

where $\mathcal{V}$ is the vocabulary. While Greedy is computationally efficient, its myopic nature may lead to suboptimal overall sequences since only the locally optimal choice is considered at each step.

Beam Search (BS) (Freitag and Al-Onaizan, 2017) addresses the limitations of Greedy by keeping track of the top-k highest-scoring partial sequences (beams) at each time step. At step t, each beam $\boldsymbol{y}_{<t}^{(i)}$ is extended with every possible next token $y \in \mathcal{V}$, producing candidates scored by the cumulative log-probability:

$$\text{score}(\boldsymbol{y}_{1:t}^{(i)}) = \sum_{\tau=1}^{t} \log P\big(y_\tau \mid \boldsymbol{y}_{<\tau}^{(i)}, \boldsymbol{x}\big). \tag{3}$$

Only the top-k candidates are retained as beams for the next time step, trading off exploration and efficiency. BS often yields higher-quality sequences than Greedy but can still suffer from low diversity and search errors when k is small. In this study, we tuned beam size among 3, 5, and 7.

Diverse Beam Search (DBS) (Vijayakumar et al., 2018) augments classical beam search with an explicit diversity prior. All k hypotheses are partitioned into G groups of equal size k/G. At every decoding step, the algorithm first ranks candidates inside each group and then retains the top-k/G sequences per group, rather than the global top-k. The score assigned to a partial sequence $(\boldsymbol{y}_{<t}, y)$ belonging to group g is

$$\begin{aligned}\text{score}\big(\boldsymbol{y}_{<t}, y\big) = \log P\big(\boldsymbol{y}_{<t}, y, \big|, \boldsymbol{x}\big) \\ - \lambda \sum_{g'<g} \Delta\big((\boldsymbol{y}_{<t}, y), \mathcal{B}_t^{g'}\big),\end{aligned} \tag{4}$$

where $\mathcal{B}_t^{g'}$ denotes the beam of group g' at time t and $\Delta(\cdot,\cdot)$ is a similarity measure (e.g., n-gram overlap). DBS encourages beams to explore different regions of the search space, improving output variety. In this study, we tuned beam size k and group size G among (3, 3), (6, 3), (9, 3), (6, 6), and (12, 6).

Contrastive Search (CS) (Su et al., 2022; Su and Collier, 2023) assumes that the language model (LM) embeds tokens in an approximately isotropic space. Given the context $(\boldsymbol{x}, \boldsymbol{y}_{<t})$, it selects the next token by jointly maximizing likelihood and dissimilarity to the preceding hidden states:

$$\begin{aligned}y_t = \underset{y \in \mathcal{V}^k}{\operatorname{argmax}} \Big[(1-\alpha) P(y \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) \\ - \alpha \max_{1 \le j \le t-1} s\big(h_y, h_{x_j}\big)\Big],\end{aligned} \tag{5}$$

where $\mathcal{V}^k$ is the top-k candidate set, h are hidden states, and s is usually the cosine similarity. The presence of the second term causes the language model to avoid tokens that are too similar to previous ones, reducing degeneration. We select the best α among 0.2, 0.4, and 0.6.

Contrastive Decoding (CD) (Li et al., 2023) similarly incorporates a contrastive penalty but directly modifies token-level logits by using an amateur language model. For each candidate token y:

$$\text{score}\big(\boldsymbol{y}_{<t}, y\big) = (1-\beta)\boldsymbol{z}_y^e - \beta \boldsymbol{z}_y^a, \tag{6}$$

where $\boldsymbol{z}_y^e$ and $\boldsymbol{z}_y^a$ are logits in the expert language model and the amateur language model, respectively. In addition, CD introduces the following vocabulary constraints to penalize scores by taking into account the grammatical ability and commonsense of the amateur language model:

$$\begin{aligned}\mathcal{V}^{head}\big(\boldsymbol{x}, \boldsymbol{y}_{<t}\big) = \{y \in \mathcal{V} : P^e(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \\ > \alpha \max P^a(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x})\},\end{aligned} \tag{7}$$

where P^e and P^a are softmax probability in the expert language model and the amateur language model, respectively. We set $\alpha = 0.1$, and search $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Frustratingly Simple Decoding (FSD) (Yang et al., 2024) contrasts an LM (P^{base}) with an on-the-fly anti-LM (P^{anti}) estimated from the current prefix to penalize the repetition. Two instantiations exist: an n-gram model (FSD) and a vectorized model (FSD-vec).

The selection rule is

$$P^{\text{FSD}}(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) = (1-\alpha)P^{\text{base}}(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) - \alpha P^{\text{anti}}(y \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \tag{8}$$

evaluated over $\mathcal{V}^n$, the top-n tokens under P^{base}. We tuned $n \in \{3, 5\}$ and $\alpha \in \{0.3, 0.5, 0.7\}$.

Decoding by Contrastive Layers (DoLa) (Chuang et al., 2024) is a decoding strategy designed to enhance the factuality of language models. This method derives a more factual next-token distribution by contrasting the standard next-token prediction obtained from the model's final layer with a prediction from an earlier, or "premature," layer. Specifically, DoLa utilizes the difference between the logits of the final layer and those of a premature layer to adjust the distribution, thereby encouraging the selection of higher-confidence words.

The selection of the premature layer is dynamic and employs the Jensen-Shannon Divergence (JSD) as a metric to measure the distance between next-token probability distributions. From a set of candidate premature layers, the one exhibiting the largest JSD with the final layer's probability distribution is chosen. This approach aims to identify a layer that contains significantly different information compared to the final layer, thereby emphasizing their contrast. For Llama2-7B and Llama3-8B series, we selected the premature layers from [0, 16) and [16, 32).

Self-Logits Evolution Decoding (SLED) (Zhang et al., 2024a) improves the factuality of LLM outputs by evolving the model's logits during decoding to dynamically adjust the token selection process. SLED achieves this by first contrasting the logits from the model's final layer with those from selected earlier, "premature," layers to unearth potential factual inconsistencies or underexpressed knowledge. It then employs an approximate gradient-based approach, where this identified latent knowledge guides a "self-evolution" or refinement process of the output probability distribution. This iterative adjustment aims to steer the generation towards more factually accurate tokens, effectively improving truthfulness while maintaining fluency and incurring negligible latency. Consequently, SLED helps LLMs produce more reliable and factually sound text by better aligning their outputs with their inherent knowledge.

In SLED, the main hyperparameters are the top n tokens compared to the logit and the evolution rate α in the logit evolution. We search $n \in \{5, 10\}$ and $\alpha \in \{0.1, 1.0, 5.0\}$.

Task	Dataset	N
QA	TriviaQA (Joshi et al., 2017)	17,210
TS	XSum (Narayan et al., 2018)	11,334
MT	WMT19 (De-En) (Foundation, 2019)	2,998
CG	HumanEval (Chen et al., 2021)	164

Table 5: Dataset statistics.

B Details of Datasets

Dataset statistics are in Table 5.

TriviaQA (Joshi et al., 2017) is a large-scale reading-comprehension dataset including question–answer pairs authored independently of evidence documents. Each question is paired with supporting context drawn from both Wikipedia and diverse web sources, enabling evaluation of open-domain and extractive QA systems.

XSum (Narayan et al., 2018) is a large-scale, single-document abstractive summarization dataset consisting of BBC news articles paired with professionally written, single-sentence summaries.

WMT19 (Foundation, 2019) refers to the training and evaluation data released for the 2019 Workshop on Machine Translation shared task, which is designed to benchmark neural machine translation systems. It includes distinct development and test sets to measure translation for news.

HumanEval (Chen et al., 2021) is a collection of 164 programming problems, each paired with a reference implementation and a suite of unit tests.

C Instruction Templates

The instruction templates for each task are listed from Figure 3 to Figure 6.

```
# Question: {question}

# Answer:
```

Figure 3: The prompt for QA.

```
Article: {text}

Summarize the above article in 1 sentence.
```

Figure 4: The prompt for TS.

```
Translate the following sentence from German to English.
{text}
```

Figure 5: The prompt for MT.

```
Please complete the remaining Python function code based on the following docstring content.
{text}
```

Figure 6: The prompt for CG.

Model	Method	TriviaQA	XSum		WMT19			HumanEval
		RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Llama2-7B-Chat	Greedy	11.36	15.08	17.43	17.44	66.62	77.68	34.76
	BS	12.16	17.82	18.88	15.31	62.17	78.38	29.88
	DBS	10.91	17.81	18.57	15.50	62.56	78.40	31.10
	CS	10.45	17.79	17.40	18.88	68.26	77.69	37.80
	CD	5.03	14.95	17.65	12.04	57.12	62.33	15.85
	FSD	4.98	16.97	18.01	10.49	54.72	78.10	34.76
	FSD-vec	3.37	17.06	18.40	11.29	54.68	78.27	17.07
	DoLa	11.60	17.86	17.89	17.28	66.00	77.80	36.59
	SLED	10.91	–	–	14.22	63.85	78.65	46.95
Llama3-8B-RLHF	Greedy	5.75	31.19	76.71	83.07	16.86	16.47	39.63
	BS	5.70	15.02	52.13	84.35	13.92	21.19	18.90
	DBS	5.61	15.28	52.68	84.47	14.03	21.10	25.00
	CS	5.88	30.31	75.73	83.17	16.83	16.74	42.68
	CD	5.09	6.90	54.54	57.97	17.24	18.09	28.05
	DoLa	5.86	35.03	80.93	83.64	17.33	16.71	33.54
Llama3-8B-SFT	Greedy	53.00	20.60	10.37	40.16	86.04	83.05	48.17
	BS	17.14	20.07	12.73	18.99	56.97	85.19	7.93
	DBS	18.11	20.11	12.53	18.90	56.94	84.98	20.12
	CS	52.50	20.46	10.55	40.11	86.01	82.78	47.56
	CD	33.16	20.99	12.38	7.14	56.40	59.47	23.78
	DoLa	55.69	21.13	10.60	41.28	86.16	83.65	49.39

Table 6: Quality scores for every task and generation metric pair in Llama2-7B-Chat, Llama3-8B-RLHF and Llama3-8B-SFT.

D Quality Scores

Results for each quality scores are listed in Table 6.

E Additional Results on Qwen2.5 Series

Table 7 and Table 8 present the UE performance achieved with Qwen2.5-7B-Instruct (Qwen et al., 2025)[12] and Qwen2.5-14B-Instruct,[13] respectively. For CD, we used Qwen2.5-0.5B-Instruct[14] as the amateur model.

F Additional Results on Llama3-13B-Chat

Table 9 presents the UE performance achieved with Llama2-13B-Chat.[15]

G Experiments on Stochastic Decoding Strategies

To succinctly evaluate the uncertainty impact of the stochastic decoding strategies omitted from our comprehensive experiments in Section 2, we experimented on Temperature Sampling ($T \in \{0.8, 1.0, 1.2\}$) and Top-p Sampling (Holtzman et al., 2020) ($p = 0.9$). The results in Table 10 show that UE performance remains nearly identical to Greedy, suggesting that introducing stochasticity confers little reliability benefit.

H Additional Results on Advanced UE Method

We combine TokenSAR, a variant of Shifting Attention to Relevance (Duan et al., 2024), with each decoding strategy and show the results in Table 11. PRR$_{AlignScore}$ scores from Greedy and CS, and FSD-TokenSAR in MT setting outperform MSP, while the rest degrade. Existing benchmarking (Vashurin et al., 2024) that comprehensively investigated UE performance has shown that simple MSP is superior, and these results are consistent with those of the previous study.

I Details of Implementation

We used a single NVIDIA A100 40GB for all experiments. Decoding strategies have been implemented with reference to Hugging Face Transformers (Wolf et al., 2020) and official implementations.[16][17][18] Quality metrics, uncertainty metrics, and uncertainty estimation methods have been implemented with reference to LM-polygraph (Fadeeva et al., 2023).

J Settings of Hyperparameters

The optimal hyperparameters for each decoding strategy across different datasets and models are listed from Table 12 to Table 16.

[12] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[13] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[14] https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

[15] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[16] https://github.com/XiangLi1999/ContrastiveDecoding

[17] https://github.com/LHRYANG/FSD

[18] https://github.com/JayZhang42/SLED/

Method	MSP							MTE						
	TriviaQA	XSum		WMT19			HumanEval		TriviaQA	XSum		WMT19			HumanEval	
	RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1	Mean PRR	RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1	Mean PRR
Greedy	67.56	10.77	-0.32	35.33	56.03	28.55	-9.63	26.90	60.90	8.00	5.11	43.56	59.06	30.34	-9.40	28.22
BS	62.90	10.29	-2.54	32.31	52.00	25.41	0.36	25.82	56.84	5.99	3.26	40.43	53.98	28.06	0.84	27.06
DBS	63.80	4.42	-10.46	5.24	29.51	14.28	-9.27	13.93	39.29	6.87	5.34	20.08	31.59	10.32	0.64	16.30
CS	67.59	11.08	0.03	35.36	56.58	28.38	-8.15	27.27	60.93	8.22	5.38	43.60	59.56	31.08	-5.86	28.99
CD	-21.87	-23.64	4.11	17.27	-32.21	-19.28	-9.90	-12.22	48.36	11.33	3.84	29.67	0.27	8.81	-9.84	13.21
DoLa	64.58	9.46	3.58	34.63	54.37	24.08	-8.01	26.10	58.34	7.09	5.09	38.10	52.95	23.65	-10.56	24.95

Table 7: PRRs for every task and generation metric pair in Qwen2.5-7B-Instruct.

Method	MSP						MTE					
	TriviaQA	WMT19			HumanEval		TriviaQA	WMT19			HumanEval	
	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR
Greedy	71.44	41.37	67.48	27.83	2.25	42.07	66.81	47.28	65.93	31.46	6.31	43.56
BS	65.55	37.89	62.51	23.18	9.18	39.66	62.65	43.19	58.79	27.19	13.16	41.00
DBS	68.04	13.62	48.34	10.64	-7.73	26.58	42.15	19.45	31.45	12.08	-7.94	19.44
CS	71.63	41.60	67.85	27.35	5.18	42.72	66.90	47.53	66.37	30.97	4.69	43.29
CD	-9.90	17.63	21.30	-19.60	8.98	3.68	18.89	22.28	8.87	13.14	11.95	15.03
DoLa	65.00	39.96	65.36	25.84	13.31	41.89	58.94	41.14	62.21	26.52	13.17	40.40

Table 8: PRRs for every task and generation metric pair in Qwen2.5-14B-Instruct.

Method	MSP						MTE					
	TriviaQA	WMT19			HumanEval		TriviaQA	WMT19			HumanEval	
	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR
Greedy	57.08	32.92	49.87	29.15	-17.94	30.22	39.44	45.92	46.89	31.25	-16.58	29.38
BS	56.27	43.8	62.42	30.28	-16.8	35.19	32.22	33.37	26.25	21.47	-16.44	19.37
DBS	53.92	-44.82	60.03	18.71	4.57	18.48	8.52	-34.45	-66.69	-17.64	4.29	-21.19
CS	57.85	41.55	60.64	29.33	15.89	41.05	38.0	43.59	45.38	31.09	4.65	32.54
CD	6.43	42.49	56.25	59.3	-20.76	28.74	11.99	50.68	62.55	65.63	-25.72	33.03
FSD	20.56	56.52	62.52	70.69	-13.01	39.46	20.78	57.71	63.04	70.52	-11.55	40.1
FSD-vec	19.86	56.15	64.03	70.82	-14.51	39.27	20.23	57.36	64.55	70.78	-13.03	39.98
DoLa	55.64	44.0	61.33	31.38	-20.25	34.42	30.92	32.92	23.44	23.95	-20.77	18.09

Table 9: PRRs for every task and generation metric pair in Llama2-13B-Chat.

Method	MSP						MTE					
	TriviaQA	WMT19			HumanEval		TriviaQA	WMT19			HumanEval	
	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR	RougeL	BLEU	Comet	AlignScore	Pass@1	Mean PRR
Greedy	62.97	38.74	46.48	19.02	-11.03	31.24	49.13	31.24	25.03	21.69	-13.49	22.72
Temperature	63.73	38.45	46.19	19.35	-13.92	30.76	51.22	30.15	23.69	21.34	-14.00	22.48
Top-p	61.82	38.65	46.79	19.00	-11.34	30.98	51.16	30.92	26.01	21.27	-13.32	23.21

Table 10: PRRs for every task and generation metric pair in Llama2-7B-Chat with Temperature Sampling and Top-p sampling.

Method	TriviaQA	WMT19		
	RougeL	BLEU	Comet	AlignScore
Greedy-MSP	62.97	38.74	46.48	19.02
CS-MSP	63.73	36.94	41.99	19.58
FSD-MSP	33.84	31.82	14.04	8.15
DoLa-MSP	61.15	38.78	49.15	16.74
Greedy-TokenSAR	51.77	29.26	24.11	21.85
CS-TokenSAR	51.45	30.83	26.93	23.45
FSD-TokenSAR	-2.14	58.36	59.46	63.68
DoLa-TokenSAR	50.19	28.16	23.40	15.32

Table 11: PRRs for every task and generation metric pair in Llama2-7B-Chat with TokenSAR (Duan et al., 2024).

Model	Method	TriviaQA	XSum		WMT19			HumanEval
		RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Llama2-7B-Chat	Greedy	-	-	-	-	-	-	-
	BS	3	3	7	3	3	5	3
	DBS	9_3	9_3	6_3	6_3	3_3	3_3	9_3
	CS	0.2	0.2	0.2	0.6	0.6	0.6	0.6
	CD	0.7	0.5	0.5	0.3	0.1	0.1	0.5
	FSD	5_0.7	3_0.3	3_0.5	3_0.3	3_0.3	5_0.5	5_0.3
	FSD-vec	5_0.5	3_0.3	5_0.5	3_0.5	3_0.5	3_0.5	3_0.5
	DoLa	[16, 32)	[0, 16)	[16, 32)	[0, 16)	[0, 16)	[0, 16)	[16, 32)
	SLED	5.0_5	1.0_5	1.0_5	0.1_5	5.0_10	0.1_5	5.0_5
Llama3-8B-RLHF	Greedy	-	-	-	-	-	-	-
	BS	7	3	5	3	3	3	3
	DBS	9_3	3_3	9_3	12_6	3_3	6_3	3_3
	CS	0.6	0.2	0.4	0.2	0.2	0.2	0.6
	CD	0.3	0.5	0.5	0.5	0.5	0.5	0.1
	DoLa	[16, 32)	[16, 32)	[16, 32)	[0, 16)	[16, 32)	[16, 32)	[0, 16)
Zephyr-7B-β	Greedy	-	-	-	-	-	-	-
	BS	7	7	5	3	3	3	5
	DBS	9_3	9_3	9_3	9_3	9_3	3_3	3_3
	CS	0.2	0.2	0.4	0.4	0.6	0.4	0.4
	FSD	3_0.3	3_0.3	3_0.3	5_0.3	5_0.5	5_0.5	5_0.5
	FSD-vec	3_0.3	3_0.3	3_0.3	5_0.3	5_0.3	5_0.5	3_0.5
	DoLa	[0, 16)	[0, 16)	[0, 16)	[0, 16)	[0, 16)	[16, 32)	[0, 16)

Table 12: Optimal hyperparameters in Table 1.

Method	TriviaQA	XSum		WMT19			HumanEval
	RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Greedy	-	-	-	-	-	-	-
BS	7	3	5	3	3	5	3
DBS	9_3	3_3	9_3	12_6	3_3	6_3	3_3
CS	0.6	0.2	0.4	0.2	0.2	0.2	0.6
CD	0.3	0.5	0.5	0.5	0.5	0.5	0.1
DoLa	[16, 32)	[16, 32)	[16, 32)	[0, 16)	[16, 32)	[16, 32)	[0, 16)

Table 13: Optimal hyperparameters in Llama3-8B-SFT.

Method	TriviaQA	XSum		WMT19			HumanEval
	RougeL	RougeL	AlignScore	BLEU	Comet	AlignScore	Pass@1
Greedy	-	-	-	-	-	-	-
BS	7	3	5	5	5	5	3
DBS	9_3	9_3	9_3	6_3	9_3	6_3	9_3
CS	0.6	0.4	0.2	0.4	0.2	0.2	0.6
CD	0.1	0.5	0.7	0.7	0.3	0.5	0.5
FSD	5_0.7	5_0.5	5_0.5	3_0.5	5_0.5	5_0.3	3_0.7
FSD-vec	3_0.7	5_0.5	3_0.5	5_0.3	3_0.7	5_0.3	5_0.5
DoLa	[0, 16)	[0, 16)	[16, 32)	[0, 16)	[0, 16)	[16, 32)	[0, 16)

Table 14: Optimal hyperparameters in Table 7.

Method	TriviaQA	WMT19			HumanEval
	RougeL	BLEU	Comet	AlignScore	Pass@1
Greedy	-	-	-	-	-
BS	7	5	5	7	5
DBS	9_3	9_3	9_3	9_3	6_6
CS	0.2	0.2	0.2	0.2	0.4
CD	0.1	0.1	0.3	0.3	0.5
DoLa	[16, 32)	[0, 16)	[0, 16)	[0, 16)	[16, 32)

Table 15: Optimal hyperparameters in Table 8.

Method	TriviaQA	WMT19			HumanEval
	RougeL	BLEU	Comet	AlignScore	Pass@1
Greedy	-	-	-	-	-
BS	3	3	3	3	5
DBS	6_3	6_3	6_3	6_6	6_6
CS	0.2	0.6	0.6	0.4	0.6
CD	0.1	0.1	0.1	0.1	0.1
DoLa	[0, 16)	[0, 16)	[0, 16)	[0, 16)	[16, 32)

Table 16: Optimal hyperparameters in Table 9.