Breaking the Trade-Off Between Faithfulness and Expressiveness for Large Language Models

Chenxu Yang[1,2*] , **Qingyi Si**[3*], **Lanrui Wang**[1,2], **Zheng Lin**[1,2†]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Huawei Technologies Co., Ltd.
{yangchenxu,linzheng}@iie.ac.cn; siqingyi@huawei.com

Abstract

Grounding responses in external knowledge represents an effective strategy for mitigating hallucinations in Large Language Models (LLMs). However, current LLMs struggle to seamlessly integrate knowledge while simultaneously maintaining *faithfulness* (or *fidelity*) and *expressiveness*, capabilities that humans naturally possess. This limitation results in outputs that either lack support from external knowledge, thereby compromising faithfulness, or appear overly verbose and unnatural, thus sacrificing expressiveness. In this work, to break the trade-off between faithfulness and expressiveness, we propose **Co**llaborative **De**coding (**CoDe**), a novel approach that dynamically integrates output probabilities generated with and without external knowledge. This integration is guided by distribution divergence and model confidence, enabling the selective activation of relevant and reliable expressions from the model's internal parameters. Furthermore, we introduce a knowledge-aware reranking mechanism that prevents over-reliance on prior parametric knowledge while ensuring proper utilization of provided external information. Through comprehensive experiments, our plug-and-play CoDe framework demonstrates superior performance in enhancing faithfulness without compromising expressiveness across diverse LLMs and evaluation metrics, validating both its effectiveness and generalizability.

1 Introduction

Although large language models (LLMs) have demonstrated remarkable performance across diverse tasks in recent studies (Bai et al. 2023; OpenAI 2023a,b), they remain susceptible to hallucination, producing content that appears plausible yet lacks factual accuracy (Ji et al. 2023; Huang et al. 2025). Research indicates that this phenomenon arises from fundamental limitations in LLMs, including constrained knowledge boundaries (Ren et al. 2023), insufficient coverage of long-tail knowledge (Kandpal et al. 2023), and outdated parametric knowledge. These inherent constraints significantly hinder the practical deployment of LLMs. To address these challenges, augmenting LLMs with external knowledge through incorporation into model inputs has emerged as a promising solution, demonstrating substantial improvements in the factual accuracy of generated content. The Retrieval-Augmented Generation (RAG) paradigm, in particular, has gained widespread adoption as an effective approach to this problem.

Figure 1: Examples exhibits the trade-off between expressiveness and faithfulness in LLMs. Higher x-coordinates correspond to higher faithfulness, and higher y-coordinates correspond to better expressiveness. Examples (a), (c), and (d) are constrained by the trade-off, whereas our approach break it and generate responses like (b).

However, external-knowledge-augmented LLMs, such as those employing RAG, continue to face two fundamental challenges. First, they frequently generate content that contradicts or lacks support from provided knowledge, as shown in responses (a) and (c) of Figure 1. Second, they struggle to integrate external knowledge naturally, often producing responses with poor interactivity, dullness, and redundancy (Chen et al. 2023; Yang et al. 2023). Response (d) in Fig-

* Equal Contribution
† Zheng Lin is the corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ure 1 illustrates this limitation: the model merely echoes the provided knowledge without addressing the user's greeting, substantially diminishing conversational engagement. While existing methods address the first challenge (Zhang et al. 2024; Liang et al. 2024), they neglect or even worsen the second. An effective LLM must balance two requirements: it must generate responses grounded in the provided knowledge, a property we define as faithfulness (or fidelity), and it must leverage external knowledge creatively to produce natural, diverse, and engaging responses, which we term expressiveness. We provide detailed definitions for these two properties in Section 3.2. Previous work by Chawla et al. (2024) identified this fidelity-expressiveness conflict through input masking experiments, yet failed to propose a practical solution. We extend this analysis to decoding strategies in Section 3.3, revealing that deterministic decoding sacrifices expressiveness while stochastic decoding compromises fidelity. This comprehensive understanding enables our principled solution.

To resolve this trade-off in LLMs, we propose Collaborative Decoding (CoDe), a novel method that dynamically elicits relevant and factual natural expressions from the model's internal parameters. CoDe achieves this by integrating output probability distributions generated with and without external knowledge, guided by their distributional divergence and model confidence. Specifically, we employ Jensen-Shannon Divergence (JSD) to quantify the distribution differences and combine local confidence (maximum probability) with global uncertainty (entropy) to measure model confidence, facilitating complementary cooperation between two distributions. By enhancing expressiveness without introducing stochasticity, our approach effectively circumvents hallucinations typically associated with sampling-based methods. Additionally, we introduce a knowledge-aware reranking mechanism to prevent over-reliance on parametric knowledge at the expense of external knowledge. This mechanism reranks the top-k candidate tokens based on their alignment with external knowledge, evaluated by both semantic similarity and attention patterns, thereby ensuring faithfulness to the provided knowledge.

Our contributions are summarized as follows:

- We investigate the trade-off between expressiveness and faithfulness in external-knowledge-augmented LLMs, focusing on decoding strategies.
- We introduce CoDe, a novel method that simultaneously enhances both faithfulness and expressiveness in knowledge-grounded scenarios without requiring additional training, model, or generation budgets.
- We demonstrate CoDe's effectiveness and generalizability through comprehensive experiments, comparing against ten baseline decoding methods across six LLMs, three datasets, and nine evaluation metrics.

2 Related Work

2.1 Hallucinations in Text Generation

Hallucination refers to the generation of LLMs appears plausible but is factually incorrect (Zhang et al. 2023b). The research community has extensively investigated this phenomenon from multiple perspectives, including its underlying causes (Dziri et al. 2022b), detection methodologies (Manakul, Liusie, and Gales 2023), and mitigation strategies (Chuang et al. 2023). Retrieval-Augmented Generation (RAG) has emerged as a prominent approach for mitigating hallucinations by incorporating external knowledge. Several studies have pursued training-based solutions, constructing preference-aligned or human-annotated datasets to fine-tune models for improved fidelity (Liang et al. 2024; Zhang et al. 2024). Others have adopted Chain-of-Thought approaches (Wei et al. 2022), externalizing implicit knowledge from the backbone LLM or employing self-reflection mechanisms (Asai et al. 2024). In contrast, our CoDe method offers a lightweight solution that effectively mitigates hallucinations without requiring training, auxiliary models, or time-intensive reflections.

2.2 Generation Decoding Strategy

Decoding strategies determine next-token selection from vocabulary probability distributions, including greedy decoding, beam search, and top-k sampling. Nucleus sampling (Holtzman et al. 2020) dynamically selects tokens until reaching a cumulative probability threshold. While stochastic methods enhance diversity, they compromise semantic consistency (Su et al. 2022) and increase hallucination rates (Dziri et al. 2022a). Recent contrastive decoding methods have recently gained significant attention. Contrastive Decoding (Li et al. 2023b) maximizes expert-amateur log-probability differences for improved fluency. DoLa (Chuang et al. 2023) contrasts mature and pre-mature layer logits to reduce hallucinations. CAD (Shi et al. 2024) amplifies probability differences between context-aware and context-free outputs. VCD (Leng et al. 2023) contrasts original and distorted visual inputs in vision-language models. Unlike these error-filtering approaches, CoDe employs dynamic collaboration between distributions, simultaneously optimizing both fidelity and expressiveness rather than addressing single limitations.

3 Preliminaries

3.1 Task Formulation

We consider an LLM parameterized by θ. The model input comprises four components: a task-specific instruction $\mathcal{I}$, multi-turn dialogue history $\boldsymbol{h}$, the current user utterance $\boldsymbol{u}$, and relevant external knowledge $\boldsymbol{k} = (k_1, \ldots, k_m)$ containing m tokens. For notational convenience, we define the conversation context as $\boldsymbol{x} = [\mathcal{I}; \boldsymbol{h}; \boldsymbol{u}]$.

At each time step, the LLM generates the next token based on the input and previously generated tokens $\boldsymbol{y}_{<t}$, producing vocabulary logits:

$$\text{logit}_\theta(y_t|\cdot) = \mathcal{LLM}_\theta(\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t}). \tag{1}$$

The probability distribution is obtained via softmax transformation, and various decoding strategies select the next token y_t from the resulting distribution:

$$y \sim p_\theta(y_t|\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t}) \propto \exp \text{logit}_\theta(y_t|\cdot). \tag{2}$$

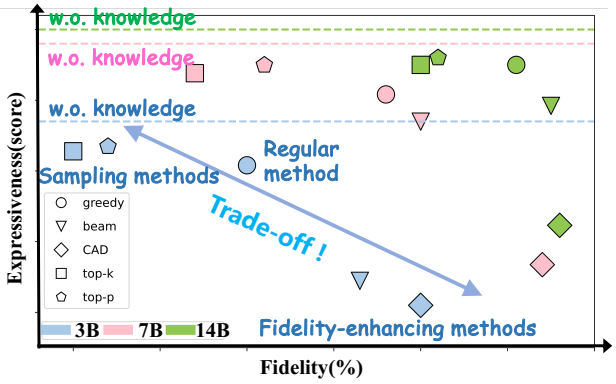


Figure 2: The trade-off between fidelity and expressiveness of current decoding strategies on Qwen2.5-chat models at different scales. The dashed line indicates the expressiveness score without referring to knowledge.

3.2 Conceptual Definitions

Faithfulness (or **fidelity**) denotes the consistency between generated responses and external knowledge without contradictions. A formal definition and distinction from **factuality** are provided in Appendix H.

Expressiveness encompasses three key dimensions: (1) *context-aware interaction*, prioritizing conversational coherence and user engagement over mere information delivery; (2) *natural knowledge integration*, extracting and seamlessly incorporating relevant information rather than copying source text; and (3) *linguistic diversity*, exhibiting varied expression patterns while avoiding formulaic language.

3.3 Pilot Observations and Insights

There remains considerable potential for improvement in expressiveness and fidelity. Integrating external knowledge into LLMs creates a fundamental tension: while improving informativeness, it often diminishes response expressiveness. As shown in Figure 1 (panels c-d) and quantified in Figure 2, LLMs tend to directly copy external knowledge rather than seamlessly incorporating it, resulting in decreased expressiveness scores. This suggests that LLMs sacrifice discourse coherence and naturalness when prioritizing faithful information transmission. Moreover, a substantial fidelity gap exists between LLM and human-generated content. Despite external knowledge access, LLMs frequently produce contradictory outputs due to flawed reasoning or conflicts with their parametric knowledge, as illustrated in Figure 1 (panels a, c). Even advanced open-source LLMs significantly underperform humans in maintaining knowledge fidelity, highlighting persistent challenges in neural knowledge grounding.

Current decoding strategies reveal a fundamental trade-off between expressiveness and knowledge fidelity. As illustrated in Figure 2, this dilemma manifests distinctly across different decoding approaches: deterministic decoding yields content with high fidelity but compromised expressiveness, whereas stochastic decoding enhances linguistic diversity at the cost of factual accuracy. Notably, this trade-off is particularly pronounced in smaller-scale models, which exhibit greater sensitivity to the choice of decoding strategy. This paper aims to break the trade-off by proposing a novel approach that achieve a win-win situation for both faithfulness and expressiveness.

4 Approach

This section presents Collaborative Decoding (CoDe), a novel method for external-knowledge-augmented LLMs comprising two key components, as illustrated in Figure 3.

4.1 Adaptive Dual-Stream Fusion

As shown in Section 3.3, models with external knowledge input tend to copy knowledge fragments, thereby diminishing expressiveness. While stochastic decoding methods like top-k (Fan, Lewis, and Dauphin 2018) and nucleus sampling (Holtzman et al. 2020) mitigates this issue, their probabilistic nature inevitably induces hallucinations. We hope to propose a deterministic approach that enhances expressiveness without sacrificing factual accuracy.

Distribution Collaboration. Inspired by contrastive decoding (Li et al. 2023b), we propose a dual-stream fusion approach that emphasizes *complementary collaboration* rather than error filtering through contrast. CoDe generates two output distributions: an expressiveness-oriented stream conditioned solely on conversation context x, and a faithfulness-oriented stream conditioned on both context x and external knowledge k. These streams are then fused to create a collaborative distribution that breaks the trade-off between expressiveness and faithfulness :

$$\begin{aligned} p_{CoDe}(y_t) = \text{softmax}[&\alpha \,\text{logit}_\theta(y_t|\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t}) \\ &+ (1-\alpha)\,\text{logit}_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})], \end{aligned} \tag{3}$$

where larger α indicates more weight on the faithfulness-oriented stream. The Equation 3 could also be written as:

$$p_{CoDe} \propto p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \left(\frac{p_\theta(y_t|\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t})}{p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})} \right)^\alpha. \tag{4}$$

In this formulation, $p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})$ represents the prior distribution based solely on the model's parametric knowledge, while $p_\theta(y_t|\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t})$ denotes the posterior distribution conditioned on external knowledge $\boldsymbol{k}$. CoDe leverages pointwise mutual information (PMI) between $\boldsymbol{k}$ and y_t to dynamically recalibrate output probabilities, amplifying tokens strongly associated with external knowledge.

Adaptive Fusion Weights α. To prevent hallucinations from low-probability tokens, we adaptively modulate α based on **model confidence** and **distribution divergence**. When internal knowledge exhibits low relevance or high uncertainty, CoDe reduces parametric reliance and prioritizes external knowledge integration.

$$\alpha = \frac{\delta \cdot \mathcal{C}_t^k}{\mathcal{C}_t^c + \delta \cdot \mathcal{C}_t^k}, \tag{5}$$

where $\mathcal{C}_t^k$ and $\mathcal{C}_t^c$ denotes the confidence of posterior and prior knowledge, δ denotes the distribution divergence.

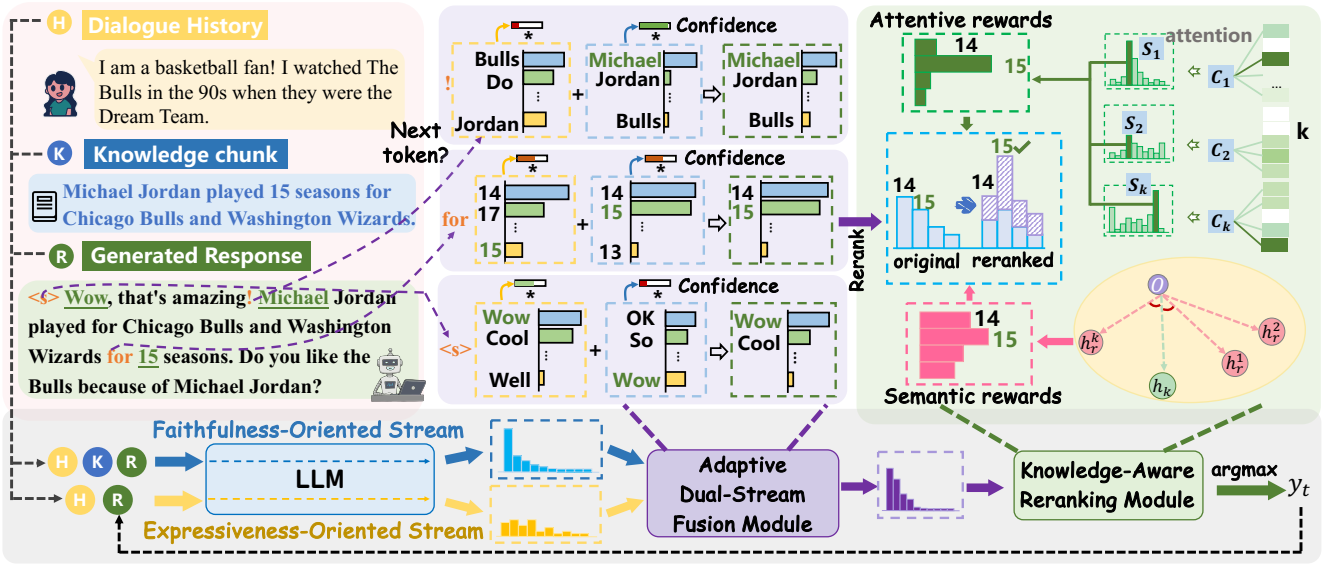


Figure 3: An overview of the CoDe method, which comprises two key components: (1) an Adaptive Dual-Stream Fusion Module that dynamically integrates internal and external knowledge by leveraging model confidence and distribution divergence, and (2) a Knowledge-Aware Reranking Module that employs semantic and attentive rewards to select faithful tokens.

Recent work on LLM hallucination determine when to trust LLMs based on uncertainty (Manakul, Liusie, and Gales 2023; Huang et al. 2023; Duan et al. 2023), we adopt the uncertainty-based confidence framework of Zhang et al. (2023a), quantifying factual confidence through local confidence p_{max} (maximum token probability) and global uncertainty $\mathcal{H}_t$ (distribution entropy):

$$p_{max} = \max_{y_t \in \mathcal{V}} p(y_t),$$
$$\mathcal{H}_t = -\sum_{y_t \in \mathcal{V}} p(y_t) * \log_2(p(y_t)). \tag{6}$$

We then synthesize p_{max} and $\mathcal{H}_t$ using the geometric mean function, deriving the confidence score $\mathcal{C}_t$ as follows:

$$\mathcal{C}_t = \sqrt[2]{\frac{p_{max}}{\mathcal{H}_t + \eta}}, \tag{7}$$

where η is a small constant prevents value overflow.

When the prior and posterior distributions diverge significantly, this signals a conflict between internal and external knowledge, prompting us to reduce the prior weight and prioritize external information. Conversely, when the distributions align closely, indicating consistent knowledge representations, we increase the prior weight to leverage pretrained knowledge for enhanced expressiveness. To implement this adaptive mechanism, we introduce a dynamic parameter δ in the design of α:

$$\delta = \gamma \cdot \exp(\text{JSD}\left(p_c(y_t) \| p_k(y_t)\right)), \tag{8}$$

where $\text{JSD}(\cdot,\cdot)$ denotes the Jensen-Shannon Divergence, and γ is a scale factor.

4.2 Knowledge-Aware Reranking

To prevent the model from being overly confident in its prior parameter knowledge and thereby ignoring external knowledge, we introduce a knowledge-aware reranking mechanism that further refines CoDe's output distribution:

$$\hat{p}_{CoDe}(y_t) = \text{topK}\Big\{(1-\beta)\, p_{CoDe}(y_t) + \frac{\beta}{2}\Big[\max_{k_i \in \boldsymbol{k}}\{\text{sim}(h_{y_t}, h_{k_i})\} + \max_{k_j \in \boldsymbol{k}}\{\text{att}(y_t, k_j)\}\Big]\Big\}, \tag{9}$$

where β controls the fidelity amplification strength, h represents hidden states, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\text{att}(y_t, k_j)$ represents the max-pooled attention weight between token y_t and knowledge element k_j across all layers and heads. The knowledge-aware reranking mechanism ensures fidelity through two complementary rewards: (1) semantic reward, which favors tokens with high cosine similarity to external knowledge tokens, and (2) attentive reward, which prioritizes tokens exhibiting stronger attention to knowledge segments. As illustrated in Figure 3, when internal and external knowledge conflict (e.g., the model's "14 seasons" versus the correct "15 seasons" for Jordan), this mechanism amplifies external knowledge awareness, enabling accurate token selection ($14 \rightarrow 15$).

The final token y_t is selected from the top-K candidates based on the combined score of fidelity and expressiveness:

$$y_t = \arg\max \hat{p}_{CoDe}(y_t). \tag{10}$$

5 Experiments

5.1 Experimental Setup

Datasets and Models. e evaluated CoDe on three information-seeking dialogue datasets—FAITHDIAL (Dziri

Method	FAITHDIAL							HalluDial						
	Expressiveness			Faithfulness			Avg.	Expressiveness			Faithfulness			Avg.
	DIV	COH	CRE	F-Critic	H-Judge	K-BP		DIV	COH	CRE	F-Critic	H-Judge	K-BP	
Greedy	31.4	57.3	30.0	28.5	86.9	60.9	49.2	36.9	64.6	30.1	30.2	87.8	61.2	51.8
Beam	30.8	**57.6**	25.6	31.3	89.3	64.9	49.9	36.1	64.4	24.3	31.5	89.0	65.7	51.8
CS	33.9	55.4	30.0	30.9	83.2	58.7	48.7	37.5	64.6	30.2	30.4	88.7	60.9	52.0
FECS	32.8	56.8	28.0	31.6	88.1	63.5	50.1	39.4	64.4	30.4	31.0	89.9	64.3	53.2
top-k	**36.2**	57.2	**34.5**	21.8	75.3	56.8	47.0	40.7	63.8	**36.0**	21.1	73.0	56.4	48.5
Nucleus	35.6	57.2	34.3	23.4	79.7	57.4	47.9	39.9	64.1	34.6	25.3	79.0	57.8	50.1
F-Nucleus	34.1	57.3	32.9	24.3	82.0	58.6	48.2	38.5	64.4	32.3	25.7	82.4	59.1	50.4
CD	35.0	55.9	31.3	22.6	76.2	57.0	46.3	38.4	62.9	30.5	24.1	78.9	57.3	48.7
DoLa	32.8	56.2	32.3	31.4	87.3	61.2	50.2	39.0	64.0	33.6	32.2	89.1	60.4	53.0
CAD	29.2	52.8	21.7	32.1	90.4	**67.0**	48.9	35.4	59.8	22.3	33.6	**90.4**	67.3	51.5
CoDe	35.6	**57.6**	29.9	**32.4**	**90.8**	**67.0**	**52.2**	**40.9**	**64.9**	29.8	**34.3**	**90.4**	**67.5**	**54.6**

Table 1: Automatic evaluation results on the FAITHDIAL and HalluDial dataset (Llama2-7B-chat). The best results are highlighted with **bold**. The second-best results are highlighted with underline. Avg. denotes the average across all metrics.

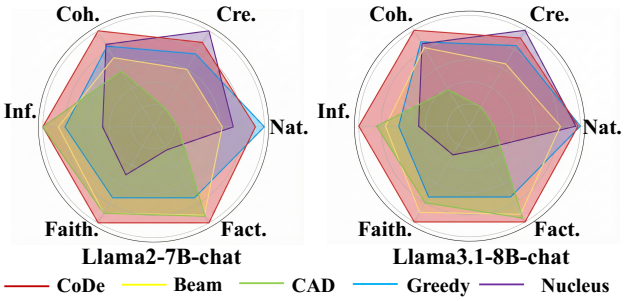


Figure 4: LLM-based evaluation results on the FAITHDIAL dataset (Llama2-7B-chat).

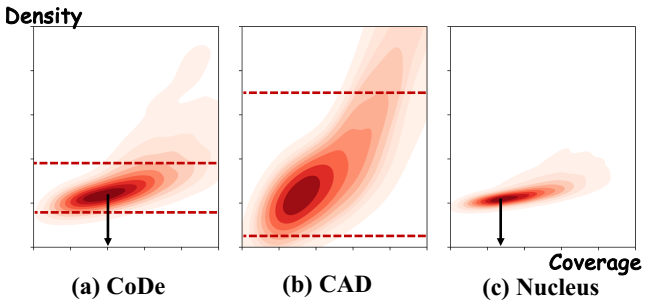


Figure 5: Knowledge utilization patterns across CoDe, CAD, and Nucleus decoding methods. Bottom-right concentration indicates superior performance.

et al. 2022a), HalluDial (Luo et al. 2024), and WoW (Dinan et al. 2018)—which provide dialogue contexts with external knowledge for response generation. Additionally, we tested on three non-conversational benchmarks: Natural Questions (Kwiatkowski et al. 2019), NQ-SWAP (Longpre et al. 2022), and HalluEval (Li et al. 2023a), demonstrating CoDe's effectiveness in faithfulness-only scenarios. We evaluated six LLMs across different scales and architectures: Llama2-7B-chat, Llama-3.1-8B-chat, Mistral-7B-Instruct-v0.2 (Jiang et al. 2023), and Qwen-2.5 series (3B, 7B, 14B) (Qwen et al. 2025).

Baselines. We choose ten decoding methods as the baselines. **Search Methods:** Greedy Decoding (Greedy), Beam Search (Beam), Contrastive Search (CS) (Su et al. 2022), and FECS (Chen et al. 2023). **Stochastic Methods:** Top-k Sampling (Fan, Lewis, and Dauphin 2018), Nucleus Sampling (Nulceus) (Holtzman et al. 2020), and Factual-Nucleus Sampling (F-Nucleus) (Lee et al. 2023). **Contrastive Methods:** Contrastive Decoding (CD) (Li et al. 2023b), DoLa (Chuang et al. 2023), and Context-Aware Decoding (CAD) (Shi et al. 2024).

5.2 Experimental Results

Automatic Evaluation We conducted comprehensive automated evaluation using 9 metrics across 3 dimensions:

Faithfulness. We employed three metrics: K-BP (BERT-Precision between knowledge and response) (Chen et al. 2023), F-Critic (average entailment score using FaithCritic NLI model) (Dziri et al. 2022a), and H-Judge (faithfulness ratio assessed by HalluJudge LLM) (Luo et al. 2024).

Expressiveness. We assessed diversity (DIV), context coherence (COH), and creative knowledge utilization (CRE). DIV measures lexical diversity via geometric mean of Distinct-n (n=1,2,3,4) (Li et al. 2016). COH quantifies context-response alignment through cosine similarity of sentence embeddings (Su et al. 2022; Li et al. 2023b). CRE evaluates non-extractive knowledge use the COVERAGE divided by the square root of DENSITY (Grusky, Naaman, and Artzi 2020).

Quality. Overall quality was measured using standard overlap-based metrics: BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE (Lin 2004).

Results. As shown in Tables 1, CoDe consistently outperforms all ten baseline methods across three faithfulness metrics on all datasets. Our approach also achieves top-2 performance in diversity and relevance metrics. Notably, the CRE scores indicate that CoDe reduces direct knowledge copying compared to other fidelity-enhancing methods like Beam Search and CAD. We further analyzed knowledge utilization patterns, as shown in Figure 5. CoDe exhibits lower density

Model	Method	Expressiveness			Faithfulness			Quality			Avg.
		DIV	COH	CRE	F-Critic	H-Judge	K-BP	BLEU-2/4	METEOR	ROUGE-L	
Mistral-7B-Instruct-v0.2	**greedy**	34.2	59.5	41.3	21.8	89.8	59.9	15.0/6.7	19.6	25.2	37.3
	top-k	35.2	59.5	**46.8**	16.8	87.0	57.4	14.3/6.1	18.9	23.7	36.6
	CAD	33.1	58.3	35.2	23.9	91.2	62.5	15.0/6.6	20.4	**25.3**	37.1
	CoDe	**35.4**	**59.9**	38.7	**24.3**	**91.3**	**62.7**	**15.5/6.9**	**20.6**	25.0	**38.0**
Llama-3.1-8B-chat	**greedy**	34.4	53.4	34.2	46.6	92.2	67.7	21.6/10.4	22.2	**31.0**	41.4
	top-k	**36.0**	53.2	**35.1**	42.9	91.5	62.9	19.8/9.4	20.6	28.8	40.0
	CAD	29.7	50.5	23.8	49.7	93.3	**72.5**	21.0/9.8	23.0	30.3	40.4
	CoDe	35.7	**54.5**	33.8	**50.2**	**94.0**	71.7	**21.9/10.8**	**23.5**	**30.8**	**42.7**
Qwen-2.5-3B-chat	**greedy**	37.7	52.4	37.6	38.7	90.5	56.2	18.6/8.5	16.2	25.8	38.2
	top-k	**40.0**	50.7	**45.9**	29.0	85.4	53.8	16.9/7.5	15.5	23.9	36.9
	CAD	34.8	48.5	31.9	39.6	91.4	61.0	19.1/8.6	17.3	26.5	37.9
	CoDe	39.4	**54.4**	37.0	**42.9**	**92.9**	**61.3**	**20.9/9.8**	**18.9**	**27.4**	**40.5**
Qwen-2.5-7B-chat	**greedy**	36.8	55.7	40.5	36.0	91.2	61.6	16.8/7.6	18.5	25.5	39.0
	top-k	37.3	54.8	**45.5**	32.8	88.7	58.0	15.2/7.0	17.6	24.4	38.1
	CAD	35.2	52.6	34.6	38.4	**92.8**	63.6	**17.8/8.0**	20.5	26.1	39.0
	CoDe	**37.7**	**55.8**	40.8	**39.6**	**92.8**	**64.8**	17.6/**8.0**	**20.6**	**26.4**	**40.4**
Qwen-2.5-14B-chat	**greedy**	37.8	**53.6**	39.3	36.6	91.9	65.4	21.7/10.3	21.6	30.1	40.8
	top-k	38.5	53.4	**43.8**	36.2	91.6	63.8	21.3/10.0	21.5	29.4	41.0
	CAD	35.1	52.8	36.5	36.4	91.9	**66.5**	22.0/10.3	21.4	30.3	40.3
	CoDe	**38.6**	**53.6**	39.6	**36.9**	**92.6**	66.2	**22.4/10.5**	**22.0**	**30.7**	**41.3**

Table 2: Automatic evaluation results compared with SoTA baselines across five LLMs on the FAITHDIAL dataset.

Method	Acc	ROUGE-L	BERT-P	Avg.
Greedy	56.3	20.4	53.8	43.5
Beam	58.1	21.6	55.7	45.1
CS	55.9	19.2	52.0	42.4
FECS	57.6	**23.0**	57.1	45.9
F-Nucleus	49.5	18.8	48.9	39.1
DoLa	56.1	20.4	53.9	43.5
CAD	57.4	**22.9**	56.3	45.5
CoDe	**58.8**	22.4	**58.3**	**46.5**

Table 3: Evaluation results on the HalluEval (summarization) dataset (Llama2-7B-chat).

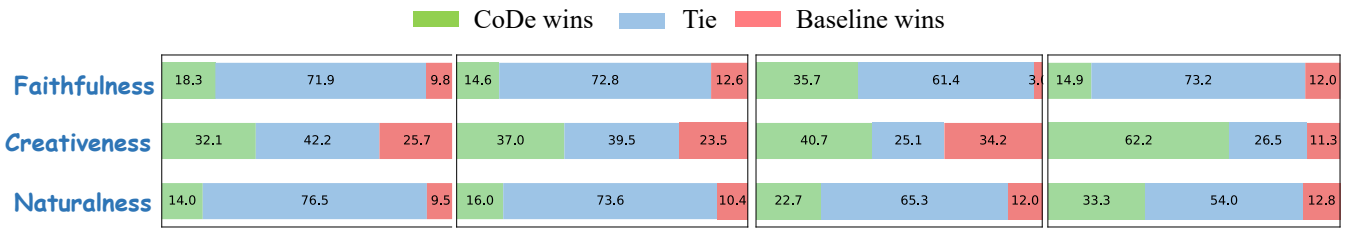
Method	NQ	NQ-SWAP	HalluEval(QA)	Avg.
Greedy	32.5	26.3	54.9	37.9
Beam	28.7	21.8	45.0	31.8
CS	30.5	22.2	52.3	35.0
FECS	34.2	29.0	57.1	40.1
F-Nucleus	24.4	18.7	49.6	30.9
DoLa	33.5	21.4	55.8	36.9
CAD	34.0	**31.9**	55.7	40.5
CoDe	**34.5**	31.6	**57.3**	**41.1**

Table 4: Accuracy (Acc) results on NQ, NQ-SWAP and HalluEval (QA) datasets (Llama2-7B-chat).

than CAD while maintaining higher coverage than sampling methods, indicating substantial token overlap with knowledge sources but minimal contiguous copying. This pattern suggests that CoDe integrates external knowledge more naturally and diversely, extracting relevant information without resorting to verbatim reproduction. Tables 7 and 8 (in Appendix) demonstrate that CoDe performs more closely to the ground-truth in traditional metrics, indicating its overall better performance. Table 2 demonstrates that CoDe significantly improves both fidelity and expressiveness across diverse model architectures and scales. On FAITHDIAL, CoDe achieves H-Judge improvements of +3.9% for Llama2-7B-chat and +2.4% for Qwen-2.5-3B-chat over greedy decoding. Remarkably, CoDe enables the 3B model to surpass larger models on multiple metrics (DIV, COH, F-Critic, and H-Judge), highlighting its efficiency in resource-constrained settings. The results in Tables 3 and 4 demonstrate that CoDe also achieves strong performance on QA and summarization benchmarks that focus solely on faithfulness, highlighting the generalizability of our decoding strategy across diverse task settings.

LLMs-based Evaluation We employed GPT-4.1 for LLM-as-a-Judge evaluation (Liu et al. 2023; Zheng et al. 2023) on 200 randomly sampled FAITHDIAL test instances. Five decoding methods were evaluated across six criteria (1-5 scale): Naturalness, Coherence, Informativeness, Creativity, Faithfulness, and Factuality, following established rating protocols (Fu et al. 2023). Figure 4 demonstrates that CoDe successfully overcomes the expressiveness-fidelity trade-off, achieving superior overall performance. While nucleus sampling and CAD show bias toward either dimension, CoDe outperforms greedy search across nearly all criteria, confirming the automated evaluation results in Table 2.

Human Evaluation To complement automated and LLM-based evaluations, we conducted human evaluation on 200 randomly selected FaithDial test samples. Five well-educated annotators compared responses from CoDe and baseline methods across three criteria: Naturalness, Creativity, and Faithfulness. As shown in Figure 6, CoDe significantly outperformed all baselines in faithfulness. For creativity, annotators preferred CoDe 1.25× over greedy search and 5.5× over CAD. For naturalness, CoDe was favored 1.5×



(a) CoDe vs. Greedy (b) CoDe vs. Beam (c) CoDe vs. Nucleus (d) CoDe vs. CAD

Figure 6: Human evaluation results on the FAITHDIAL dataset (Llama2-7B-chat). The result is statistically significant with p-value < 0.05, and Kappa (κ) falls between 0.5 and 0.7, suggesting moderate agreement.

Setup		Expressiveness			Faithfulness			Avg.
		DIV	COH	CRE	F-Critic	H-Judge	K-BP	
A	CoDe	35.2	57.6	29.9	**32.4**	**90.8**	67.0	**52.2**
B	-α	34.9	57.5	32.1	30.1	89.2	64.7	51.4
C	-EOS	34.7	56.8	27.3	32.3	**90.8**	**67.3**	51.5
D	-Sem	35.0	57.1	29.6	31.4	88.6	64.1	51.0
E	-Att	35.2	57.5	30.4	30.9	88.3	63.6	51.0
F	-KAR	**35.6**	**58.0**	**33.9**	29.1	85.5	59.3	50.2

Table 5: Ablation study on the FAITHDIAL dataset. Avg. denotes the average across all metrics.

over greedy search and 1.9× over nucleus sampling.

5.3 Ablation Study

This section presents ablation studies examining key components and hyperparameters (β, γ). We evaluated five components on FAITHDIAL using Llama2-7B-chat: Dynamic Fusion Weight (-α), Expressiveness-Oriented Stream (-EOS), Knowledge-Aware Reranking (-KAR), semantic reward (-Sem), and attentive reward (-Att). Table 5 shows that all modules are essential. Setup A (equal-weight stream fusion) degrades performance across both dimensions, emphasizing the importance of adaptive internal knowledge utilization. Removing EOS reduces expressiveness similar to other fidelity-focused baselines. Without KAR, expressiveness slightly improves but fidelity drops unacceptably. Setups D and E confirm both reward mechanisms are necessary.

5.4 Qualitative Examples

We demonstrate CoDe's ability to generate informative and engaging responses through several case studies. In Figure 8, baseline methods exhibit distinct failure modes: Greedy Search misidentifies Madonna's cover as an original work; both Greedy Search and DoLa confuse song titles with album names, generating unverifiable content; DoLa produces accurate but excessively long lists without proper discourse structure; and CAD merely copies external information without addressing user queries. In contrast, CoDe maintains natural interaction while correctly integrating external knowledge.

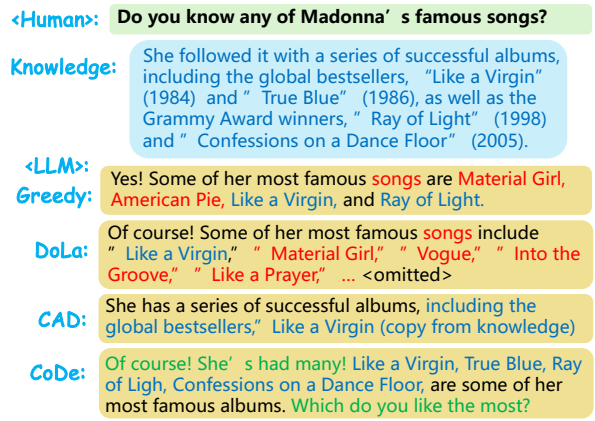


Figure 7: Case study example with unfaithful hallucinations marked in red and faithful information in blue.

6 Conclusion

In this paper, we systematically investigated the inherent trade-off between faithfulness and expressiveness in external-knowledge-augmented LLMs. To address this limitation, we introduced Collaborative Decoding (CoDe), a plug-and-play method that dynamically integrates parametric and external knowledge through adaptive dual-stream fusion and knowledge-aware reranking. Extensive experiments across six LLMs and multiple benchmarks demonstrate that CoDe successfully overcomes the faithfulness-expressiveness trade-off. This work opens new avenues for developing decoding strategies that leverage the complementary strengths of internal and external knowledge sources, ultimately advancing the capabilities of LLM assistants in real-world applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472419, 62472420).

References

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Cri-

tique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. arXiv:2309.16609.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Chawla, K.; Rashkin, H.; Tomar, G. S.; and Reitter, D. 2024. Investigating Content Planning for Navigating Trade-offs in Knowledge-Grounded Dialogue. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2316–2335. St. Julian's, Malta: Association for Computational Linguistics.

Chen, W.-L.; Wu, C.-K.; Chen, H.-H.; and Chen, C.-C. 2023. Fidelity-Enriched Contrastive Search: Reconciling the Faithfulness-Diversity Trade-Off in Text Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 843–851. Singapore: Association for Computational Linguistics.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. arXiv:2309.03883.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *CoRR*, abs/1811.01241.

Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; and Xu, K. 2023. Shifting Attention to Relevance: Towards the Uncertainty Estimation of Large Language Models. arXiv:2307.01379.

Dziri, N.; Kamalloo, E.; Milton, S.; Zaiane, O.; Yu, M.; Ponti, E. M.; and Reddy, S. 2022a. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10: 1473–1490.

Dziri, N.; Milton, S.; Yu, M.; Zaiane, O.; and Reddy, S. 2022b. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? arXiv:2204.07931.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898. Melbourne, Australia: Association for Computational Linguistics.

Fu, J.; Ng, S.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. *CoRR*, abs/2302.04166.

Grusky, M.; Naaman, M.; and Artzi, Y. 2020. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. arXiv:1804.11283.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. arXiv:1904.09751.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2).

Huang, Y.; Song, J.; Wang, Z.; Zhao, S.; Chen, H.; Juefei-Xu, F.; and Ma, L. 2023. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. arXiv:2307.10236.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; and Raffel, C. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 452–466.

Lee, N.; Ping, W.; Xu, P.; Patwary, M.; Fung, P.; Shoeybi, M.; and Catanzaro, B. 2023. Factuality Enhanced Language Models for Open-Ended Text Generation. arXiv:2206.04624.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13872–13882.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023a. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv:2305.11747.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 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*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023b. Contrastive Decoding: Open-ended Text Generation as Optimization. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12286–12312. Toronto, Canada: Association for Computational Linguistics.

Liang, Y.; Song, Z.; Wang, H.; and Zhang, J. 2024. Learning to Trust Your Feelings: Leveraging Self-awareness in LLMs for Hallucination Mitigation. In Yu, W.; Shi, W.; Yasunaga, M.; Jiang, M.; Zhu, C.; Hajishirzi, H.; Zettlemoyer, L.; and Zhang, Z., eds., *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, 44–58. Bangkok, Thailand: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *CoRR*, abs/2303.16634.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2022. Entity-Based Knowledge Conflicts in Question Answering. arXiv:2109.05052.

Luo, W.; Shen, T.; Li, W.; Peng, G.; Xuan, R.; Wang, H.; and Yang, X. 2024. HalluDial: A Large-Scale Benchmark for Automatic Dialogue-Level Hallucination Evaluation. arXiv:2406.07070.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896.

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; rong Wen, J.; and Wang, H. 2023. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. In *International Conference on Computational Linguistics*.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 783–791. Mexico City, Mexico: Association for Computational Linguistics.

Su, Y.; Lan, T.; Wang, Y.; Yogatama, D.; Kong, L.; and Collier, N. 2022. A Contrastive Framework for Neural Text Generation. arXiv:2202.06417.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Yang, C.; Lin, Z.; Wang, L.; Tian, C.; Pang, L.; Li, J.; Ho, Q.; Cao, Y.; and Wang, W. 2023. Multi-level Adaptive Contrastive Learning for Knowledge Internalization in Dialogue Generation. arXiv:2310.08943.

Zhang, T.; Qiu, L.; Guo, Q.; Deng, C.; Zhang, Y.; Zhang, Z.; Zhou, C.; Wang, X.; and Fu, L. 2023a. Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 915–932. Singapore: Association for Computational Linguistics.

Zhang, X.; Peng, B.; Tian, Y.; Zhou, J.; Jin, L.; Song, L.; Mi, H.; and Meng, H. 2024. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1946–1965. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023b. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.