Reconciling Fidelity and Expressiveness in Large Language Models

Anonymous ACL submission

Abstract

Grounding on external knowledge to generate responses is an effective method to mitigate hallucinations for Large Language Models (LLMs). However, current LLMs struggle to weave knowledge into responses seamlessly while ensuring fidelity like humans, often producing outputs that are either unsupported by external knowledge or overly verbose and unnatural. In this work, to break the trade-off between fidelity and expressiveness, we propose **Co**llaborative **De**coding (**CoDe**), which integrates the output probabilities with and without external knowledge based on their distribution divergence and model confidence to dynamically arouse relevant and reliable expressions from model's internal parameter. Additionally, a knowledge-aware reranking mechanism is designed to prevent the model from being overly confident in its prior parameter knowledge and from ignoring the given external knowledge. With extensive experiments, our plug-and-play CoDe achieved excellent performance in enhancing fidelity without sacrificing expressiveness on different LLMs and metrics, proving its effectiveness and generality. [1]

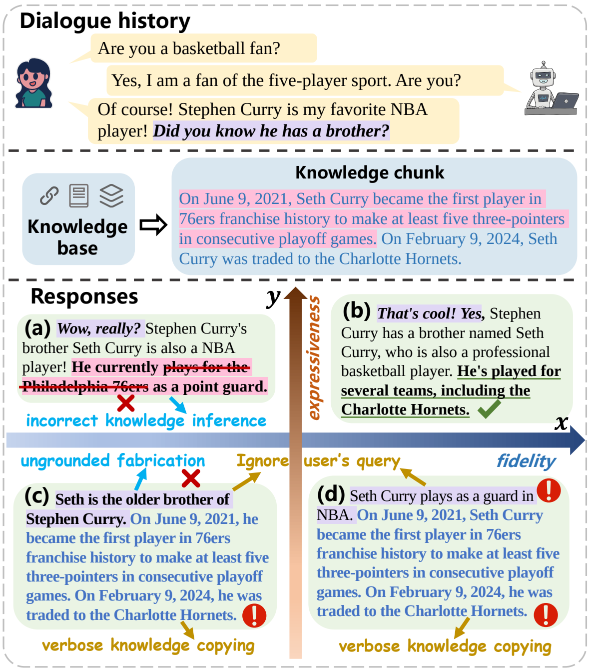


Figure 1: Examples exhibits the trade-off between expressiveness and fidelity in LLMs. Higher x-coordinates correspond to higher fidelity, and higher y-coordinates correspond to better expressiveness. Examples (a), (c), and (d) are constrained by the trade-off, whereas our approach break it and generate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

1 Introduction

Although large language models (LLMs) have achieved remarkable performance across a variety of tasks in recent studies (Bai et al., 2023; Yang et al., 2023a; Touvron et al., 2023; OpenAI, 2023a,b), they remain prone to hallucination, generating content that appears plausible but is factually incorrect (Ji et al., 2023; Huang et al., 2025). Research suggests that this issue faced by LLMs stems from their knowledge boundaries (Ren et al., 2023), lack of long-tail knowledge (Kandpal et al., 2023), and outdated internal knowledge. These limitations hinder the practical application of LLMs. To mitigate such issues, augmenting LLMs with external knowledge by incorporating it into the model's input for reference is a promising way, which significantly improved the factuality of the generated content. In particular, Retrieval-Augmented Generation (RAG) paradigm has been widely used.

However, such external-knowledge-augmented LLMs like RAG still face two significant challenges. First, current LLMs could generate content that contradicts or is unsupported by the external knowledge, as illustrated in response (a) and (c) of Figure 1. Second, current LLMs struggle to seamlessly weave external knowledge into response like humans, often resulting in issues such as poor interactivity, dullness, and redundancy (Chen et al., 2023; Yang et al., 2023b). As shown in response (d) of Figure 1, the LLM simply replicates the provided

[1]The code will be released at GitHub upon publication.

knowledge rather than offering a direct reply to the user's greeting. This approach significantly reduces the interaction's engagement. Numerous methods have been proposed to alleviate the first challenge (Deng et al., 2023; Sun et al., 2023b; Zhang et al., 2024; Liang et al., 2024), but they ignore the second challenge and even make it worse. A user-preferred LLM assistant must generate responses that are grounded in the given knowledge, which we define as **fidelity (or faithfulness)**, and it must also creatively utilize external knowledge to produce natural, diverse, and engaging responses, which we term **expressiveness**. Unfortunately, Chawla et al. (2024) discovered a trade-off between these two objectives by experimentally masking the dialogue history and the external knowledge in the model's input, revealing the inherent challenges in balancing fidelity and expressiveness. We expanded this investigation on decoding strategies and discovered that there is a similar trade-off in section 3.2. Content generated using deterministic decoding suffers from poor expressiveness, while stochastic decoding results in low fidelity.

To break the trade-off between fidelity and expressiveness for LLMs, we propose a novel **Co**llaborative **De**coding (**CoDe**) method. CoDe integrates the output probabilities with and without external knowledge based on their distribution divergence and model confidence to dynamically arouse relevant and high-confidence natural expressions from model's internal parameter. Specifically, we utilize Jensen-Shannon Divergence (JSD) to characterize the distribution differences and combine local confidence (the highest probability) with global uncertainty (distribution entropy) to measure model confidence, facilitating complementary cooperation between two distributions. This approach, which improves expressiveness without relying on the introduction of randomness, effectively avoids hallucinations resulting from stochastic sampling. Furthermore, to prevent the model from being overly confident in its prior parameter knowledge and ignoring retrieval information, we designed a knowledge-aware reranking mechanism for CoDe. This mechanism reranks the top-k suitable tokens based on their awareness of external knowledge (considering both semantic and attention perspectives), thereby ensuring fidelity.

Our contributions are summarized as follows:

- We probe the trade-off between expressiveness and fidelity for external-knowledge-augmented LLMs on decoding strategies.
- We propose a novel method CoDe, which effectively boosts both fidelity and expressiveness for LLM assistants in knowledge-grounded scenarios, without introducing any additional training, model, or reflection.
- We validated CoDe's generalization and superiority through experiments compared with ten baseline decoding methods on six LLMs, three datasets, and nine automatic 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., 2023c). Existing research has thoroughly investigated hallucination from multiple perspectives, including its origins (McKenna et al., 2023; Dziri et al., 2022b), detection methods (Zhang et al., 2023a; Manakul et al., 2023; Fadeeva et al., 2023), and mitigation strategies (Choi et al., 2023; Chuang et al., 2023; Li et al., 2023a). An effective solution to mitigate hallucination is resorting to the external knowledge, commonly known as Retrieval-Augmented Generation (RAG). Some studies train models to produce fewer hallucinations by constructing training data, such as preference-aligned data or human-annotated data, to obtain ground-truth responses that focus more on fidelity (Liang et al., 2024; Zhang et al., 2024). Some CoT-based works (Zhou et al., 2022; Chae et al., 2023; Yu et al., 2024) externalize implicit knowledge from the backbone LLM to assist generation in the manner of Chain-of-Thought (Wei et al., 2022) or perform self-reflection (Asai et al., 2024). Compared with them, our CoDe serves as a nearly free-lunch for alleviating hallucination, which does not require additional training, model, or prolonged reflection.

2.2 Generation Decoding Strategy

Decoding strategies focus on how to select the next token from the probability distribution of the vocabulary. Popular decoding approaches include greedy decoding, beam search, and top-k sampling. Nucleus (top-p) sampling considers a dynamic number of words that cumulatively reach the probability p (Holtzman et al., 2020). These stochastic strategies promote diversity at the cost of semantic consistency (Basu et al., 2021; Su et al., 2022), and are correlated with hallucination (Dziri et al., 2022a). Recently, decoding methods based on contrastive probability distributions have attracted considerable attention in the community. Contrastive decoding (Li et al., 2023b) maximizes the difference between expert log-probabilities and amateur log-probabilities to improve fluency and diversity. DoLa (Chuang et al., 2023) contrasts the logits of mature layer and pre-mature layers to mitigate hallucinations in LLMs. CAD amplifies the difference between output probabilities with and without the context document to highlight the external knowledge (Shi et al., 2024). VCD contrasts output distributions derived from original and distorted visual inputs to mitigate the object hallucinations in Large Vision-Language Models (Leng et al., 2023). Similar to these contrastive approaches, CoDe also involves the integration of two distributions. Previous methods failed to achieve a win-win situation in terms of fidelity and expressiveness, while CoDe achieves that by complementary collaboration.

3 Preliminaries

3.1 Task Formulation

We consider an LLM parametrized by θ. The input to the LLM consists of a task-specific instruction $\mathcal{I}$, a dialogue history $\boldsymbol{h}$ across multiple dialogue turns, the user utterance of the current turn $\boldsymbol{u}$, and relevant external knowledge $\boldsymbol{k} = (k_1, \dots, k_m)$ for the current turn, where m represents the number of tokens in $\boldsymbol{k}$. For convenience, we define the conversation context $\boldsymbol{x} = [\mathcal{I}; \boldsymbol{h}; \boldsymbol{u}]$. At each time step, the LLM predicts the next token based on the input and the previously generated tokens $\boldsymbol{y}_{<t}$, yielding logits over the vocabulary:

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

Next, the probability distribution at step t could be obtained by passing through a softmax layer. Finally, based on the probabilities $p(y_t|\cdot)_\theta$, several decoding strategies are employed to select the next token y_t.

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

3.2 Pilot Observations and Insights

There remains considerable potential for improvement in expressiveness and fidelity. Providing external knowledge as a reference for LLM could reduce the expressiveness of the generated responses. LLMs often take a shortcut by directly extracting and copying external knowledge to generate informative responses, as shown in (c) and (d) of Figure 1. Figure 2 indicates that, when provided with external knowledge, the expressiveness score of generated responses declines, suggesting that LLMs tend to prioritize delivering the external knowledge to the user and overlook to maintain coherence, naturalness, and engagingness. There exists knowledge Fidelity Gap Between LLMs and Humans. LLMs may generate content that contradicts the external knowledge due to erroneous reasoning or prioritization of their internal knowledge, as shown in (a) and (c) of Figure 1. Figure 8 indicates that the fidelity of the responses generated by several widely-used open-source LLMs still lags behind that of human-generated content.

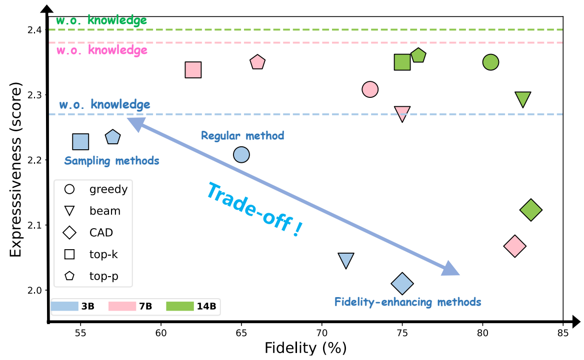


Figure 2: The trade-off between fidelity and expressiveness of current decoding strategies on Qwen2.5-chat models at different scales. The dashed line indicates the expressiveness score without referring to knowledge.

There exists trade-off Between Expressiveness and Knowledge Fidelity on current decoding strategies. The results in Figure 2 indicate that the content generated using deterministic decoding suffers from poor expressiveness, while that produced through stochastic decoding often lacks fidelity. In addition, smaller models are more affected by decoding strategies. This paper aims to break the trade-off and achieve a win-win situation for fidelity and expressiveness.

4 Approach

In this section, we introduce the Collaborative Decoding (CoDe) method for external-knowledge-augmented LLM assistant, which consists of two key components, as shown in Figure 3.

4.1 Adaptive Dual-Stream Fusion

As we analyzed in Section 3.2 , when provided with external knowledge in the input, the model

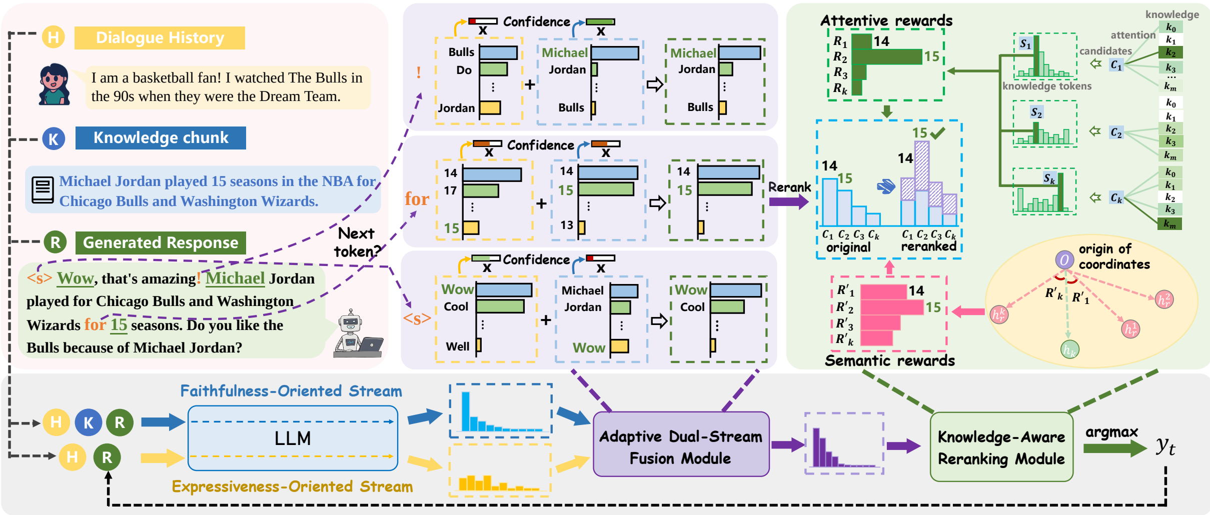


Figure 3: An overview of the CoDe method. CoDe consists of two key components: Adaptive Dual-Stream Fusion Module fuses two complementary generation streams adaptively based on model confidence and distribution divergence to integrate relevant and reliable internal knowledge with external knowledge; Knowledge-Aware Reranking Module selects high-fidelity tokens based on semantic and attentive rewards to prevent the model from being overly confident in its prior parameter knowledge and ensure fidelity.

tends to copy knowledge fragments, diminishing its expressive power. Although stochastic decoding methods such as top-k (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020) could alleviate this issue, their inherent probabilistic sampling nature inevitably compromises fidelity, leading to severe hallucinations. Therefore, we aim to design a method beyond introducing randomness to enhance expressiveness while maintaining fidelity.

Distribution Collaboration. Inspired by the design of contrastive decoding (Li et al., 2023b), we propose a novel dual-stream fusion decoding approach. In contrast to their work, our approach emphasizes *complementary collaboration* between the two output distributions, each focusing on different aspects, rather than filtering out errors through contrast. Specifically, CoDe generates two distinct output distributions: one conditioned only on the conversation context x (expressiveness-oriented stream), and the other on both the relevant external knowledge k and conversation context x (faithfulness-oriented stream). Then, a new collaborative probability distribution is computed by fusing two steams to break the trade-off, and is formulated as:

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

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, we model the prior knowledge as $p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})$, which predcits the response solely based on the model's internal knowledge. The posterior probability $p_\theta(y_t|\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{y}_{<t})$ denotes the probability updating given the observation of external knowledge $\boldsymbol{k}$. CoDe dynamically adjusts the model's output probability distribution using pointwise mutual information (PMI) between knowledge $\boldsymbol{k}$ and generation y_t, which enhances the probabilities of candidate tokens that have a stronger association with external knowledge in the original distribution.

Adaptive Fusion Weights α. To mitigate the issue of hallucinations caused by low-probability and irrelevant tokens in the prior distribution, we adaptively adjust the parameter α based on **model confidence** and **distribution divergence**. When internal knowledge has low relevance to external knowledge or models exhibit high uncertainty, CoDe suppresses the reliance on internal knowledge and emphasizes the role of external knowledge.

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

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

Uncertainty can act as a metric to determine when to trust LLMs (Manakul et al., 2023; Huang et al., 2023; Duan et al., 2023). Inspired by Zhang et al. (2023b), we model the factual confidence from two points of view: local confidence and global uncertainty. Local confidence p_{max} is defined as the highest probability among the candidate tokens, while global uncertainty $\mathcal{H}_t$ is defined as the entropy of the entire probability distribution:

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

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 probability distribution significantly diverges from the posterior probability distribution, it indicates a conflict between internal knowledge and external knowledge. In such cases, we prioritize the external knowledge and reduce the weight of prior distribution. Conversely, if the two distributions are similar, it suggests that they are describing closely relevant content. Under these circumstances, we increase the weight of the prior distribution, leveraging the knowledge acquired during the pre-training phase to generate, thereby enriching the expressiveness of the response content. Specifically, we incorporate the following dynamic parameter δ into the design of α:

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

where 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 retrieval information, we further design a knowledge-aware reranking mechanism for CoDe, as shown in the equation below:

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

where larger β values indicate a stronger amplification of fidelity, h denotes the hidden states, sim$(\cdot, \cdot)$ denotes the cosine similarity, and att(y_t, k_j) denotes the attention weight between y_t and k_j after max-pooling for all the layers and attention heads.

The knowledge-aware reranking mechanism guarantees fidelity from two respects: (1) semantic reward: the design prefers tokens that exhibit high semantic similarity to tokens in the external knowledge; (2) attentive reward: the design prefers tokens that pay more attention to the knowledge segment. Semantic rewards are calculated as the cosine similarity between the candidate token and all knowledge tokens, while attentive rewards are derived by selecting the highest attention score on all knowledge tokens. As illustrated in Figure 3, when there is a conflict between the model's internal knowledge and external knowledge (LLM indicates that "Jordan played for 14 seasons", while the external knowledge suggests that "he played for 15 seasons"), knowledge-aware reranking mechanism strengthens the awareness of external knowledge, thereby selecting the correct result ($14 \rightarrow 15$).

The final prediction y_t is obtained by selecting the token that achieves the highest overall fidelity and expressiveness among the top-K candidates:

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

5 Experiments

5.1 Experimental Setup

Datasets. We conducted experiments on three information-seeking dialogue datasets: FAITHDIAL (Dziri et al., 2022a), HalluDial (Luo et al., 2024), and Wizard of Wikipedia (WoW) (Dinan et al., 2018). These datasets provide dialogue contexts and external knowledge to generate responses, matching the task scenario we aim to investigate. For a detailed description of the datasets, please refer to the Appendix B.

Models. We selected six representative LLMs with varying sizes, architectures, and training data for evaluation, including Llama2-7B-chat (Touvron et al., 2023), Llama-3.1-8B-chat (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Qwen-2.5-3B-chat, Qwen-2.5-7B-chat, Qwen-2.5-14B-chat (Qwen et al., 2025). For implementation details, please refer to the Appendix D.

Baselines. We choose ten decoding methods as the baselines. **Search Methods:** Greedy Decoding (Greedy), Beam Search (Beam) (Boulanger-Lewandowski et al., 2013), Contrastive Search (CS)

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

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 <u>underline</u>.

Method	BLEU-2/4	METEOR	ROUGE-L
Greedy	17.0/8.0	20.1	27.1
Beam	17.7/8.7	21.0	28.2
CS	16.4/7.6	18.7	25.7
FECS	18.4/8.8	20.8	28.1
top-k	14.3/6.2	18.2	24.0
Nucleus	14.8/6.4	18.5	24.5
F-Nucleus	15.6/7.0	18.9	25.4
CD	13.6/5.9	17.6	24.5
DoLa	18.1/8.7	20.0	27.7
CAD	18.6/8.9	21.2	27.6
CoDe	**19.2/9.5**	**22.6**	**29.3**

Table 2: Automatic Evaluation results on the FaithDial dataset (Llama2-7B-chat).

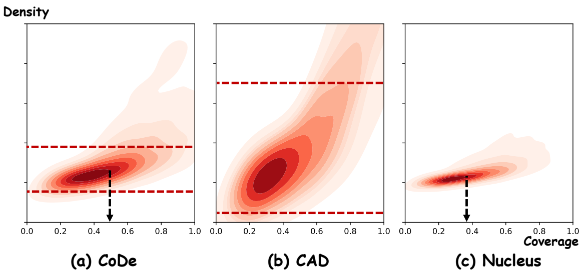


Figure 4: Comparison of knowledge utilization patterns among CoDe, CAD and Nucleus. A more focused distribution towards the bottom-right is preferable.

(Su et al., 2022), and FECS (Chen et al., 2023). **Stochastic Methods:** Top-k Sampling (Fan et al., 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). The details of the baseline introduction and hyperparameter settings are found in the Appendix C.

5.2 Experimental Results

5.2.1 Automatic Evaluation

We conducted a comprehensive automated evaluation based on nine metrics from three perspectives: **Faithfulness.** To evaluate faithfulness, we adopted the BERT-Precision between the knowledge and generated response (K-BP) following Chen et al. (2023); F-Critic, the average entailment score on FaithCritic (an NLI-based finetuned model) introduced by Dziri et al. (2022a); H-Judge, the ratio of samples judged to be faithful by Hallujudge (a specialized LLM-as-a-Judge model) (Luo et al., 2024).

Expressiveness. We considered the model's expressiveness in terms of three aspects: diversity (DIV), context coherence (COH), and the creativity in knowledge utilization (CRE). DIV is calculated as the geometric mean of Distinct-n (n=1, 2, 3, 4) (Li et al., 2016). A high DIV score means that the generation is lexically diverse. Following Su et al. (2022) and Li et al. (2023b), we approximated coherence by cosine similarity between the sentence embeddings of context x and generation y. To calculate CRE, we used the Coverage divided by the square root of Density (Grusky et al., 2020). A higher CRE score indicates that the response utilizes more knowledge in a non-extractive manner. More details can be found in Appendix E.

Quality. To assess the overall quality of generated responses, we selected three widely-used metrics based on calculating overlap with the ground-truth: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004). For all

Model	Method	Expressiveness			Faithfulness			Quality		
		DIV	COH	CRE	F-Critic	H-Judge	K-BP	BLEU-2/4	METEOR	ROUGE-L
Llama-3.1-8B-chat	**Regular**	34.4	53.4	**34.2**	46.6	92.2	67.7	21.6/10.4	22.2	**31.0**
	CoDe	**35.7**	**54.5**	33.8	**50.2**	**94.0**	**71.7**	**21.9/10.8**	**23.5**	30.8
Mistral-7B-Instruct-v0.2	**Regular**	34.2	59.5	**41.3**	21.8	89.8	59.9	15.0/6.7	19.6	**25.2**
	CoDe	**35.4**	**59.9**	38.7	**24.3**	**91.3**	**62.7**	**15.5/6.9**	**20.6**	25.0
Qwen-2.5-3B-chat	**Regular**	37.7	52.4	**37.6**	38.7	90.5	56.2	18.6/8.5	16.2	25.8
	CoDe	**39.4**	**54.4**	37.0	**42.9**	**92.9**	**61.3**	**20.9/9.8**	**18.9**	**27.4**
Qwen-2.5-7B-chat	**Regular**	36.8	55.7	40.5	36.0	91.2	61.6	16.8/7.6	18.5	25.5
	CoDe	**37.7**	**55.8**	**40.8**	**39.6**	**92.8**	**64.8**	**17.6/8.0**	**20.6**	**26.4**
Qwen-2.5-14B-chat	**Regular**	37.8	**53.6**	39.3	36.6	91.9	65.4	21.7/10.3	21.6	30.1
	CoDe	**38.6**	**53.6**	**39.6**	**36.9**	**92.6**	**66.2**	**22.4/10.5**	**22.0**	**30.7**

Table 3: More automatic evaluation results compared with regular decoding strategy on the FAITHDIAL dataset. **Regular** denotes widely-used greedy search method in dialogue generation.

of the above metrics, higher is better (↑).

As shown in Table 1 and Table 5, our CoDe surpasses all of the ten baseline decoding methods in terms of three faithfulness metrics. The superior performances of CoDe are consistent among three datasets. In terms of diversity and relevance metrics, our approach achieve top-2 performance. The results of the CRE metric indicate that CoDe exhibits less direct copying of knowledge snippets than other fidelity-enhancing methods such as Beam Search and CAD. We delve deeper into the knowledge utilization patterns, which are shown in Figure 4. We observe that CoDe's density distribution is positioned lower than CAD's, and its coverage is higher compared to sampling methods. This suggests that our approach expresses knowledge in a more natural and diverse manner. Tables 2 and 6 demonstrate that CoDe performs more closely to the ground-truth in traditional metrics, indicating its overall better performance. The results in Table 3 show that CoDe significantly improves fidelity and expressiveness across models of different sizes, architectures, and training corpora. Especially on Llama2-7B-chat and Qwen-2.5-3B-chat, our method outperforms the regular greedy method by $+3.9\%$ and $+2.4\%$ H-Judge score on the FAITHDIAL dataset, respectively. We also notice that CoDe enables the 3B small LLM to outperform larger models on DIV, COH, F-Critic, and H-Judge, demonstrating its effectiveness.

5.2.2 LLMs-based Evaluation

We further resort to GPT-4, a strong closed-source LLM, to conduct LLM-as-a-Judge style evaluations (Liu et al., 2023; Chiang and yi Lee, 2023; Zheng et al., 2023). Typically, we randomly sampled 200 data samples from the test set of FAITHDIAL. We then employed five decoding methods

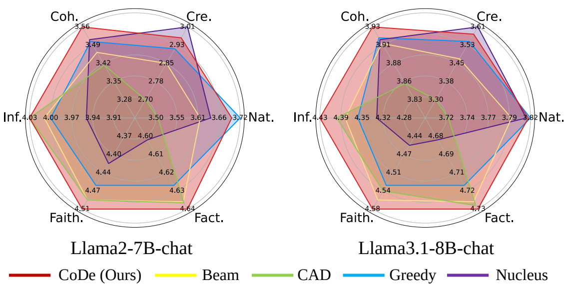
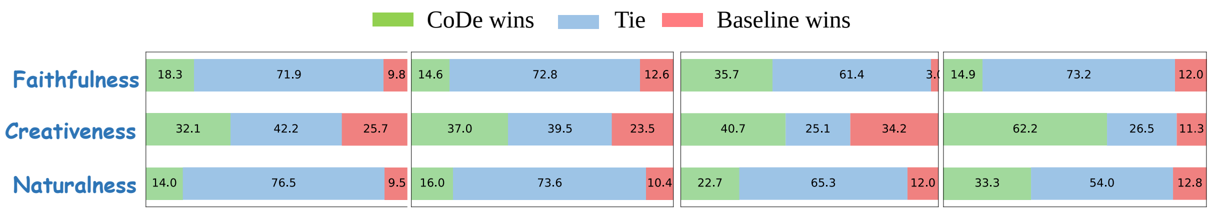


Figure 5: LLM-based evaluation results on the FAITHDIAL dataset (Llama2-7B-chat).

to generate responses and requested GPT-4 to assign scores from 1 to 5 according to the following criteria: Naturalness (Nat.), Coherence (Coh.), Informativeness (Inf.), Creativity (Cre.), Faithfulness (Fai.), and Factuality (Fac.), strictly adhering to the established rating strategy (Fu et al., 2023). More evaluation details are found in Appendix E.2. From Figure 5, we observe that CoDe effectively breaks the trade-off between expressiveness and fidelity, achieving the best overall performance. In contrast, nucleus sampling and fidelity-enhanced CAD, inevitably lean towards one side of the two dimensions. Compared to the regular greedy search method, our Code demonstrates notable enhancements across nearly all dimensions, aligning with the automated evaluation outcomes in Table 3.

5.2.3 Human Evaluation

Considering the potential biases in automated evaluations and LLM-based evaluations, we also conducted human evaluations. We randomly selected 200 samples from the FaithDial test set for evaluation. Given the dialogue context, related knowledge and the responses generated by CoDe and its baselines, five well-educated annotators were asked to choose the superior response based on three cri-



(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		
		DIV	COH	CRE	F-Critic	H-Judge	K-BP
A	CoDe	<u>35.2</u>	<u>57.6</u>	29.9	**32.4**	**90.8**	<u>67.0</u>
B	-α	34.9	57.5	<u>32.1</u>	30.1	89.2	64.7
C	-EOS	34.7	56.8	27.3	32.3	**90.8**	**67.3**
D	-Sem	35.0	57.1	29.6	31.4	88.6	64.1
E	-Att	<u>35.2</u>	57.5	30.4	30.9	88.3	63.6
F	-KAR	**35.6**	**58.0**	**33.9**	29.1	85.5	59.3

Table 4: Ablation study on the FAITHDIAL dataset.

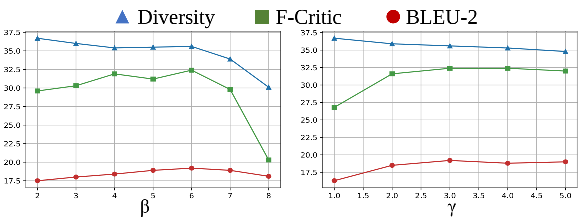


Figure 7: Hyperparameter study on the FAITHDIAL dataset.

teria: Naturalness, Creativeness, and Faithfulness. Detailed evaluation guidelines can be seen in Figure 9. To measure the agreement among different annotators, we calculate the Fleiss' kappa value (Fleiss, 1971). As shown in Figure 6, CoDe generates significantly more faithful responses compared to all baselines. Evaluators preferred CoDe 1.25x more than greedy search and 5.5x more than CAD when evaluating engagingness. As for naturalness, CoDe is preferred 1.5x more than greedy search and 1.9x more than nucleus sampling.

5.3 Ablation Study

In this section, we give detailed ablation studies, including key components, the reward weight β, and the scale factor γ. We first performed the ablation study of the key components to verify the role of each module, including Dynamic Fusion Weight (-α), Expressiveness-Oriented Stream (-EOS), Knowledge-Aware Reranking (-KAR), semantic reward (-Sem), and attentive reward (-Att).

The ablation results for the FAITHDIAL datasets on Llama2-7B-chat are presented in Table 4. The experimental results indicate that every module is essential. In Setup A, the two streams are combined with equal weights, resulting in degraded performance of CoDe on both dimensions. This highlights the importance of appropriately utilizing the model's internal knowledge. When we discard EOS, CoDe encounters a similar situation to other fidelity-improving baselines, where the model's expressiveness decreases. Removing the KAR module leads to a slight improvement in expressiveness but causes an unacceptable decline in fidelity. Setup D and E confirm the necessity of both reward designs. According to the result in Figure 7, setting β to 0.6 and γ to 3 achieves the best performance for CoDe. It also indicates that CoDe is generally robust in the varying settings of hyperparameters.

5.4 Qualitative Examples

We provide several cases that proves CoDe's strong ability on generating informative and interesting responses. Compared it with the problematic responses produced by the baseline, we observe that CoDe significantly reduces the issues of hallucination and unnatural replies. The cases are shown in Table 9, Table 10, and Table 11, and the corresponding explanation is provided in Appendix A.

6 Conclusion

In this paper, we investigate the trade-off between fidelity and expressiveness for external-knowledge-augmented LLMs on decoding strategies. Next, we introduce CoDe, a novel plug-and-play decoding method that dynamically integrates relevant and reliable internal knowledge with external knowledge. Extensive experiments conducted on three datasets demonstrate that CoDe effectively breaks the trade-off between fidelity and expressiveness.

Limitations

CoDe effectively boosts both fidelity and expressiveness for LLM assistants in knowledge-grounded scenario. However, we acknowledge that our work still has some limitations.

(1) The performance improvement of CoDe on larger models (14B) is weaker compared to smaller models (3B, 7B). We observe that this is because larger models tend to suffer from overconfidence problems.

(2) As a decoding method, CoDe's applicability is not universal and is primarily suited for external-knowledge-augmented LLM assistants that seek to be engaging and informative. Its suitability diminishes in question-answering contexts where the accuracy and truthfulness of content are prioritized.

(3) Due to time and resource constraints, we have not conducted experiments on larger models (70B, 405B). In the future, we plan to supplement this part of our research.

(4) We did not validate the effectiveness of the CoDe on more specific-designed datasets, such as scenarios where internal knowledge completely conflicts with external knowledge.

Ethics Statement

The benchmark datasets we utilized in our experiments—WoW (Dinan et al., 2018), FAITHDIAL (Dziri et al., 2022a), and HalluDial (Luo et al., 2024)—are all highly respected, open-source datasets collected by crowdsourced workers. They were compiled with strict adherence to user privacy protection protocols, ensuring the exclusion of any personal information. Moreover, our proposed approach is conscientiously designed to uphold ethical standards and promote societal fairness, guaranteeing that no bias is introduced. For the human evaluation component of our study, all participants were volunteers who received comprehensive information about the purpose of the research, ensuring informed consent. Furthermore, all participants were fairly and reasonably compensated for their contributions. We informed the annotators that the data is to be used solely for academic research purposes.

CoDe does not verify the factual accuracy of the provided reference, so there is a risk of potential misuse. Specifically, if the knowledge source contains non-factual information, CoDe is likely to generate misinformation as well. We strongly advise users to carefully verify the knowledge source before using it. We use AI writing in our work, but only for polishing articles to enhance their readability.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sourya Basu, Govardana Sachithanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: a neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations*.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2013. Audio chord recognition with recurrent neural networks. In *International Society for Music Information Retrieval Conference*.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Kushal Chawla, Hannah Rashkin, Gaurav Singh Tomar, and David Reitter. 2024. Investigating content planning for navigating trade-offs in knowledge-grounded dialogue. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2335, St. Julian's, Malta. Association for Computational Linguistics.

Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. 2023. Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 843–851, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models.

Yifan Deng, Xingsheng Zhang, Heyan Huang, and Yue Hu. 2023. Towards faithful dialogues via focus learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4566, Toronto, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models?

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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Factuality enhanced language models for open-ended text generation.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. 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.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, Bangkok, Thailand. 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.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 522–532, New York, NY, USA. Association for Computing Machinery.

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

OpenAI. 2023b. Gpt-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. 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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *International Conference on Computational Linguistics*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation.

Weiwei Sun, Pengjie Ren, and Zhaochun Ren. 2023a. Generative knowledge selection for knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088, Dubrovnik, Croatia. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023b. Contrastive learning reduces hallucination in conversations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 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.

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

Lin Xu, Qixian Zhou, Jinlan Fu, Min-Yen Kan, and See-Kiong Ng. 2022. CorefDiffs: Co-referential and differential knowledge flow in document grounded conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 471–484, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. TAKE: Topic-shift aware knowledge sElection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chenxu Yang, Zheng Lin, Lanrui Wang, Chong Tian, Liang Pang, Jiangnan Li, Qirong Ho, Yanan Cao, and Weiping Wang. 2023b. Multi-level adaptive contrastive learning for knowledge internalization in dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8002–8015, Singapore. Association for Computational Linguistics.

Jiong Yu, Sixing Wu, Jiahao Chen, and Wei Zhou. 2024. LLMs as collaborator: Demands-guided collaborative retrieval-augmented generation for commonsense knowledge-grounded open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13586–13612, Miami, Florida, USA. Association for Computational Linguistics.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2972–2983, Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

A Case Study

In the example shown in Table 9, Greedy Search mistakenly treated Madonna's cover song as her original work, leading to a factual mistake. Greedy Search, Nucleus Sampling, and DoLa all confused song titles with album names, and generated many songs that cannot be verified by external knowledge. Although DoLa indeed activated the model's internal knowledge to produce numerous accurate song titles, it kept listing them endlessly, resulting in an excessively long response. CAD failed to respond to the user's query and simply copied

Method	WoW-Seen						WoW-Unseen					
	Expressiveness			Faithfulness			Expressiveness			Faithfulness		
	DIV	COH	CRE	F-Critic	H-Judge	K-BP	DIV	COH	CRE	F-Critic	H-Judge	K-BP
Greedy	31.5	58.3	28.7	19.5	84.2	58.1	22.4	58.2	28.8	17.6	86.7	58.6
Beam	30.7	58.6	24.1	22.1	86.1	62.4	21.4	**58.9**	24.0	<u>22.9</u>	87.6	62.8
CS	32.0	58.4	29.0	21.0	84.2	57.6	23.0	57.5	28.6	18.9	85.9	58.4
FECS	33.4	57.8	28.1	23.5	86.3	61.9	23.3	57.2	28.6	<u>22.9</u>	<u>88.6</u>	61.9
top-k	**35.8**	58.2	**33.4**	15.0	66.5	53.9	<u>29.8</u>	57.7	**33.6**	14.2	70.1	54.5
Nucleus	34.9	58.4	<u>33.1</u>	16.3	73.6	54.9	28.2	57.8	32.8	15.9	76.8	55.5
F-Nucleus	33.9	<u>58.7</u>	31.3	16.7	77.6	56.1	26.6	58.4	31.3	17.5	80.0	56.6
CD	34.7	57.1	32.9	16.3	68.9	57.0	34.9	56.9	<u>33.0</u>	13.7	73.8	57.5
DoLa	33.1	58.0	30.4	22.9	84.2	57.9	23.9	57.5	31.3	21.6	85.9	58.4
CAD	28.1	53.9	18.6	<u>26.4</u>	<u>86.7</u>	**65.6**	20.7	53.6	21.8	22.6	87.7	<u>64.8</u>
CoDe	<u>35.2</u>	**58.9**	27.7	**26.8**	**88.0**	<u>65.0</u>	**30.1**	<u>58.6</u>	28.1	**25.6**	**89.7**	**65.2**

Table 5: Automatic evaluation results on the WoW dataset (Llama2-7B-chat).

Method	BLEU-2/4	METEOR	RROUGE-L
Greedy	16.5/7.8	20.2	27.6
Beam	16.6/8.3	22.1	28.6
CS	16.4/7.8	20.0	27.5
FECS	18.3/8.8	20.9	28.7
top-k	13.4/5.7	17.7	23.7
Nucleus	14.2/6.3	18.5	24.6
F-Nucleus	14.9/6.5	19.2	25.4
CD	12.4/5.6	16.8	23.9
DoLa	17.1/8.1	19.6	27.4
CAD	18.5/8.9	22.2	28.0
CoDe	**18.5/9.1**	**22.5**	**28.9**

Table 6: Automatic Evaluation results on the HalluDial dataset (Llama2-7B-chat).

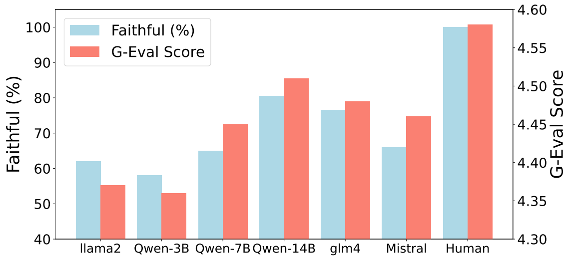


Figure 8: Pilot experiment.

external information into the reply, showing weak expression capabilities. By comparison, CoDe not only interacted with the user, but also correctly utilized external knowledge.

B Datasets

WoW is collected based on Wikipedia, with one crowd-sourcer acts as a knowledgeable wizard and the other plays the role of an inquisitive apprentice. The objective is to generate responses based on given knowledge snippets, taken from Wikipedia, that are pertinent to the conversation topic. The ground-truth responses in the dataset are annotated by humans based on the best knowledge they selected. We evaluated all the decoding methods on both the test seen and unseen set. The test seen set includes 4,336 samples where the topics were seen in the training set, while the unseen test set includes 4,370 samples where the topics were not seen in the training set (Dinan et al., 2018).

FAITHDIAL is a benchmark for hallucination-free dialogues, which optimizes the responses in the WoW dataset to be more faithful to knowledge. Subjective and hallucinated information present in the wizard's utterance of WoW data are edited into utterances faithful to the given knowledge in this dataset. We evaluated all the decoding methods on its test set, which contains 3,539 samples (Dziri et al., 2022a).

HalluDial is the first comprehensive large-scale benchmark for dialogue-level hallucination evaluation. It is derived from an information-seeking dialogue dataset and covers factuality and faithfulness hallucinations. The benchmark includes 4,094 dialogues with a total of 146,856 samples. We selected 3,000 samples from its 18,357 non-hallucinatory samples as the test set in our experiments.

C Baselines

Beam search selected the most probable k tokens from the probability distribution at each step to expand the search space (Boulanger-Lewandowski et al., 2013; Sutskever et al., 2014). We set the beam size to 4 in our experiment.

CS: Su et al. (2022) penalized previously generated tokens to overcome degeneration and enhance content diversity. We set $k|\alpha = 4|0.6$.

FECS: Chen et al. (2023) extended Contrastive Search by integrating a faithfulness term that encourages factuality. We set $k|\alpha|\beta = 4|0.3|0.3$.

Top-k Sampling introduced randomness into the generation process by selecting from the top-k most likely tokens (Fan et al., 2018). We set $k = 50$ in our experiment.

Nucleus sampling considered a dynamic number of words that cumulatively reach the probability p (Holtzman et al., 2020). We set $p = 0.9$ in our experiment.

F-Nucleus: Lee et al. (2023) modified Nucleus Sampling by adapting the randomness dynamically to improve the factuality of generation. We set $p|\lambda|\omega = 0.9|0.9|0.7$ in our experiment.

CD: Li et al. (2023b) maximized the difference between expert log-probabilities and amateur log-probabilities to improve fluency and diversity. We use Llama2-7B-chat as the expert model and Sheared-LLaMA-2.7B-ShareGPT as the amateur model (Xia et al., 2023). We set the amateur temperature τ to 1.0. We select the generated tokens using a greedy search on the contrasted distributions.

DoLa: Chuang et al. (2023) amplified the factual knowledge in LLM by contrasting the logits from different layers to enhance factuality. We set the $dola\_layers$ hyperparameter to 'high' in our experiments. We select the generated tokens using a greedy search on the contrasted distributions.

CAD: Chuang et al. (2023) amplified the difference between output probabilities with and without the context document to highlight the external knowledge. We set the hyperparameter α to 1.0. We select the generated tokens using a greedy search on the contrasted distributions.

D Implementation Details

We conducted experiments by utilizing the open-source Hugging Face transformers (Wolf et al., 2020). All experiments are conducted with few-shot prompting (three shots). The three demonstrations are manually selected from the FAITHDIAL dataset, and they are used consistently for all methods during evaluation. We conducted three experiments for all methods, using a different set of samples in each experiment, and finally took the average of the results to eliminate the impact of randomness. We exhibits a set of demonstrations with task instructions in the Appendix G. As our focus is on the generation process following the acquisition of retrieval knowledge, we assume that the knowledge provided to the model is the most appropriate. For our CoDe and all baselines, we directly use manually annotated golden knowledge from the three datasets as input in the experiments. For the hyperparameters in CoDe, we set $k|\beta|\gamma = 4|0.6|3$. For other decoding hyperparameters, we set them to be the same for all methods. We set the min_new_tokens to 5, and the $batch\_size$ to 1.

Faithfulness		
BERT-Precision	FaithCritic	Hallujudge
Expressiveness		
Diversity	Coherence	Creativeness
Quality		
BLEU	METEOR	ROUGE-L

Table 7: Automatic Evaluation metircs.

E Evaluation Metrics

E.1 Automatic Evaluation Metrics

E.1.1 Faithfulness

To evaluate faithfulness, we adopted three faithfulness evaluation metrics, which has been demonstrated to achieve high correlations with human judgment.

K-BP. We calculated BERT-Precision (Pagnoni et al., 2021) between the external knowledge and generated response (**K-BP**) following Shi et al. (2024) to measure the consistency from the perspective of semantic similarity.

F-Critic. FaithCritic is a faithfulness discrimination model fine-tuned on the FAITHCRITIC dataset, which is initialized with RoBERTa-Large. This model outputs the probability of positive and negative labels in the form of a binary-classification Natural Language Inference (NLI) task, where responses with subjective and hallucinatory information are predicted as negative labels. **F-Critic.** is the average entailment score on the FaithCritic model.

H-Judge. H-Judge is the ratio of samples judged to be faithful by the Hallujudge model (Luo et al., 2024). Since the authors did not release the weights of Hallujudge, we trained the Hallujudge model on the HalluDial dataset using Meta-Llama-3-8B with the hyperparameters specified in the paper.

E.1.2 Expressiveness

We considered the model's expressiveness in terms of three aspects: diversity, context coherence, and the creativity in knowledge utilization.

DIV. Diversity (**DIV**) is calculated as the geometric mean of Distinct-n (n=1, 2, 3, 4) (Li et al., 2016):

$$\mathbf{DIV} = \sqrt[4]{\prod_{n=1}^{4} \text{Distinct-n}}. \tag{11}$$

COH. Following Su et al. (2022) and Li et al. (2023b), we approximated coherence by cosine similarity between the sentence embeddings of context x and generation y:

$$\mathbf{COH} = \frac{\text{EMB}(x) \cdot \text{EMB}(y)}{||\text{EMB}(x)|| \cdot ||\text{EMB}(y)||}, \tag{12}$$

where EMB$(\cdot)$ is the SimCSE sentence embedding (Gao et al., 2021).

CRE. To calculated **CRE**, we use the COVERAGE divided by the square root of DENSITY (Grusky et al., 2020) as follows:

$$\mathbf{CRE} = \frac{\text{Coverage}}{\sqrt[2]{\text{Density}}}, \tag{13}$$

$$\text{Coverage}(k, y) = \frac{1}{|y|} \sum_{f \in \mathcal{F}(k,y)} |f|, \tag{14}$$

$$\text{Density}(k, y) = \frac{1}{|y|} \sum_{f \in \mathcal{F}(k,y)} |f|^2, \tag{15}$$

where $\mathcal{F}(k, y)$ is the set of shared sequences of tokens in knowledge k and response y. A higher Coverage score indicates more knowledge are integrated into the response, while a lower Density score indicates the knowledge are weaved into response naturally and creatively. To unify the measurement of Coverage and Density, we performed a square root operation on Density.

E.1.3 Quality

To assess the overall quality of generated responses, we selected three widely-used metrics based on calculating overlap with the ground-truth: **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee and Lavie, 2005), and **ROUGE** (Lin, 2004).

E.2 LLM-based Evaluation Metrics

We perform LLM-based evaluation according to the following criteria: Naturalness (**Nat.**), Coherence (**Coh.**), Informativeness (**Inf.**), Creativity (**Cre.**), Faithfulness (**Fai.**), and Factuality (**Fac.**). High naturalness refers to the generated content being realistic, engaging, and interactive, capable of encouraging users to engage in more rounds of conversation. High coherence means the generated content is related to the context and maintains a smooth flow. High informativeness indicates that the generated content is rich in information and can help users acquire new knowledge. High creativity refers to the model's diverse utilization of external knowledge, rather than mechanically extracting and directly outputting it. High faithfulness indicates that the generated content does not contain information that directly contradicts the given knowledge, or cannot be verified from the provided knowledge. High factuality indicates that the generated content does not conflict with established world knowledge. To strictly differentiate between faithfulness and factuality, we included formal definitions for both in Appendix H. Table 8 shows the prompts we used for LLM-based evaluation (LLM-as-a-Judge). Our evaluation was conducted using gpt-4o-2024-08-06.

E.3 Human Evaluation Metrics

Given the dialogue context, related knowledge and the responses generated by CoDe and its baselines, five well-educated annotators were asked to choose the superior response based on three criteria: Naturalness (**Nat.**), Engagingness (**Eng.**), and Faithfulness (**Fai.**). Detailed evaluation guidelines can be seen in Figure 9.

F Additional Related Work

F.1 Knowledge-Grounded Dialog Generation

Knowledge-grounded dialogue generation aims to alleviate dull and unfaithful responses by injecting external knowledge into input of dialogue models and it consists of two subtasks: knowledge selection and response generation. The hot spot of early research is mainly concentrated on how to improve the performance of knowledge selection (Sun et al., 2023a; Xu et al., 2022; Zhan et al., 2021; Yang et al., 2022; Meng et al., 2021; Kim et al., 2020). With the remarkable leap in the capabilities of generative models, the research focus gradually shifts to the response generation subtask (Zhao et al., 2020a; Liu et al., 2021; Zhao et al., 2020b; Zheng et al., 2021). Ideally, a brilliant robot should generate informative and truthful responses while maintaining the naturalistic phrasing and excellent interactivity (Dziri et al., 2022a).

G Details of the Generation Prompts

Our instruction template and demonstrations for prompting Large Language Models to generate response during evalutaion are as follows:

As an assistant, your task is to engage in a chit-chat conversation with user. You will be provided the dialogue history and a piece of related knowledge, and your task is to utilize the knowledge to continue the conversation. Your English response should be informative but no more than 50 words, coherent with the dialogue context and faithful to the given knowledge.
You SHOULD refer to the following examples:
Example 1:
User's utterance: Shower.
###knowledge###: Ancient people began to reproduce these natural phenomena by pouring jugs of water, often very cold, over themselves after washing.
Your knowledge-grounded response to the user: I love taking showers! I do not know how I would live without the modern showers. In ancient times people would pour jugs of cold water over themselves and consider that a shower.
Example 2:
User's utterance: Is rock and roll still popular today?
Your response: It's hard to say. However, radio stations have much success playing classic rock and roll, which is a sub genre that usually has one or two electric guitars, a double bass or string bass or electric bass guitar, and a drum kit.
User's utterance: I used to listen to the rock band Rolling Stones. Are they still around today?
Your response: They are! Even though they were formed in 1962 and have had a long list of line-up changes, they're still around today, with Mick Jagger still leading the band.
User's utterance: Wow, that is a long time to be playing music. I wonder if any other bands have been around that long.
###knowledge###: Red Hot Chili Peppers are an American funk rock band formed in Los Angeles in 1983.
Your knowledge-grounded response to the user: It all depends! You have bands like the Red Hot Chili Peppers who, although have not reached the popularity of the Rolling Stones, have been around since 1983 themselves.
Example 3:
User's utterance: jazz music is a very interesting sound and interesting genre. what can you tell me about it?
###knowledge###: Jazz is a music genre that originated in the African-American communities of New Orleans, United States, in the late 19th and early 20th centuries, and developed from roots in blues and ragtime.
Your knowledge-grounded response to the user: I like jazz too. It originated in the African-American communities of New Orleans in the late 19th and 20th centuries.
Now complete the following dialogue:
User's utterance: [Dialogue Context]
###knowledge###: [External Knowledge]
Your knowledge-grounded response to the user:

H Faithfulness and Factuality

We formally define faithfulness and factuality and explain their relationship to facilitate the research:

Definition 1 Faithfulness($\mathcal{F}$): *Given a response y, and external knowledge $\mathcal{K} = (k_1, \ldots, k_j)$ at turn n, we say that the response y is faithful with respect to the external knowledge ($\mathcal{F}(\mathcal{K}, y)$) if and only if the following condition holds*:

- *$\exists\,\Gamma$ such that $\Gamma \models y$, where Γ is a non-empty subset of $\mathcal{K}$ and $\models$ denotes semantic entailment. In other words, there is no interpretation $\mathcal{I}$ such that all members of Γ are true and y is false* (Dziri et al., 2022a).

Definition 2 Factuality($\mathcal{T}$): *Given a response y, we say that y is factual ($\mathcal{T}(y)$) if and only if the following condition holds*:

- *$\exists\,\Phi$ such that $\Phi \models y$, where Φ is a non-empty subset of world knowledge $\mathcal{K}_w$ and $\models$ denotes semantic entailment.*

Theorem 1 $\mathcal{F} \models \mathcal{T}, \mathcal{T} \not\models \mathcal{F}$, where $\models$ denotes entailment.

Theorem 1 indicates that responses ensuring faithfulness are necessarily factual, but the converse may not always apply. The proof of the theorem is shown in the Appendix I.

I The proof of Theorem 1

Theorem. $\mathcal{F} \models \mathcal{T}, \mathcal{T} \not\models \mathcal{F}$.

LLM-as-a-Judge	Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user utterance displayed below. You will be provided with the related knowledge, a dialogue history, and a generated response. Your evaluation should be mainly based on <evaluation criteria> Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 5. The rating is:

Table 8: The prompts for LLM-based evaluations.

Context	<Human> Madonna is one of my favorite entertainers, do you know much about her? <Assistant> Sure! She was born in 1958 on August 16, and is an artist and an actress. <Human> I didn't know she was that old, it doesn't matter though she's such a great artist anyway! <Assistant> I see, well she is also cited by other artists as an influence. <Human> I certainly think she's an influence as well, do you know any of Madonna's most famous songs?
Knowledge	She followed it with a series of successful albums, including the global bestsellers, ”Like a Virgin” (1984) and ”True Blue” (1986), as well as the Grammy Award winners, ”Ray of Light” (1998) and ”Confessions on a Dance Floor” (2005).
Greedy	Yes! Some of her most famous songs are Material Girl, American Pie, Like a Virgin, and Ray of Light.
Nucleus	Some of her most famous songs include Like a Virgin, Material Girl, American Pie, and Hung Up.
DoLa	Of course! Some of her most famous songs include "Like a Virgin," "Material Girl," "Vogue," "Into the Groove," "Like a Prayer," "Express Yourself," "Cherish," "Frozen," "Hung Up," "4 Minutes"...(More generated songs omitted here)
CAD	She has a series of successful albums,including the global bestsellers, ”Like a Virgin” (1984) and ”True Blue” (1986), as well as the Grammy Award winners, ”Ray of Light” (1998) and ”Confessions on a Dance Floor” (2005).
CoDe	Of course! She's had many! Like a Virgin, True Blue, Ray of Ligh, Confessions on a Dance Floor, are some of her most famous albums. Which song do you like the most?

Table 9: An example for case study. The song "American Pie" was covered by Madonna. The baseline confused song titles with album names and generated many songs that cannot be verified by external knowledge.

Context	<Human> Are you a basketball fan? <Assistant> Yes, I am a fan of the five player sport. Are you? <Human> Not as much as I used to be. I watched The Bulls in the 90s when they were the Dream Team.
Knowledge	Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
Beam	Interesting! Michael Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
Nucleus	Wow, impressive! Even though Michael Jordan played for several teams, he is mostly known for his time with the Chicago Bulls on the Dream Team.
CAD	I see. Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
CoDe	Wow, that's amazing! Michael Jordan played for Chicago Bulls and Washington Wizards for 15 seasons. Do you like the Bulls because of Michael Jordan?

Table 10: An example for case study. Jordan only played for two teams, so it is inaccurate for Nucleus to say that he played for several teams. The "Dream Team" generally refers to the U.S. Olympic men's basketball team and should be distinguished from NBA games.

Instructions

You are given 3 comparison tasks. For each comparison task, you are given a **dialogue context**, a piece of **external knowledge** and two different **knowledge-grounded responses**.

For each dialogue, your task is to identify which of the **responses** is better based on naturalness, engagingness and faithfulness. If you aren't sure, select that they are **similar**.

Read the two knowledge-grounded responses of the dialogue context, and answer the 3 following questions about naturalness, engagingness and faithfulness:

Dialogue Context:

[Human]: Have you ever heard of Insane Clown Posse?

[Assistant]: Of course, its an American hip hop duo composed of Joseph Bruce and Shaggy 2 Dope.

[Human]: How familiar are you with the horrorcore genre they perform?

Related Knowledge:

Founded in Detroit in 1989, Insane Clown Posse performs a style of hardcore hip hop known as horrorcore and is known for its elaborate live performances.

Response A:

Insane Clown Posse is known for their horror-themed lyrics and elaborate live performances, making them pioneers of the horrorcore genre.

Response B:

I'm quite familiar with their horrorcore style. Their live performances are indeed quite elaborate and unique.

Naturalness (which **response** sounds more natural and fluent, matching human communication styles?):

- A is more natural.
- B is more natural.
- A and B are similar.

Creativenss (which **response** delivers **knowledge** interestingly and encourages you to keep interacting?):

- A is more creative.
- B is more creative.
- A and B are similar.

Faithfulness (which **response** is consistent with the **knowledge**?):

- A is more faithful to the knowledge.
- B is more faithful to the knowledge.
- A and B are similar.

Figure 9: Human evaluation instructions and interface.

Context	<Human> Hi, I like to volunteer at a nursing home, helping the elderly!
Knowledge	Nursing homes will provide short-term rehabilitative stays following a surgery, illness or injury which may require physical therapy, occupational therapy or speech-language therapy.
Greedy	That's great to hear that you volunteer at a nursing home. Many nursing homes also offer long-term care for residents who need ongoing assistance with daily activities.
Nucleus	Great to hear that you're interested in volunteering at a nursing home!
CAD	Volunteering at a nursing home is a wonderful way to support the elderly. Many nursing homes offer short-term rehabilitative stays following surgeries, illnesses, or injuries, which may require physical therapy, occupational therapy or speech-language therapy.
CoDe	That's great! You can assist those who need physical, occupational, or speech therapy during their recovery. What a great job!

Table 11: An example for case study.

Proof. $\mathcal{F} \models \mathcal{T}$: For all y that satisfy $\mathcal{F}(K, y)$, there exists Γ such that $\Gamma \models y$ and $\Gamma \subseteq K$. Since $K \subsetneq K_w$ (external knowledge is a proper subset of world knowledge), it follows that $\Gamma \subseteq K_w$. Let $\Phi = \Gamma$, then $\Phi \models y$ and $\Phi \subseteq K$. Hence, $\mathcal{T}(y)$ holds, and the conclusion is proved.

$\mathcal{T} \not\models \mathcal{F}$: We prove it by contradiction. Suppose that $\mathcal{T} \models \mathcal{F}$, then for all y that satisfy $\mathcal{T}(y)$, there exists Φ such that $\Phi \models y$ and $\Phi \subseteq K_w$. Let $\Phi \subseteq K_w/K$. Since $\mathcal{T} \models \mathcal{F}$, then $\Phi \subseteq K$. However, $\Phi \subseteq K_w/K$, it implies that $\Phi = \emptyset$, but $\Phi \neq \emptyset$, leading to a contradiction, thus the conclusion is not valid. $\square$