

# Harnessing RLHF for Robust Unanswerability Recognition and Trustworthy Response Generation in LLMs

Shuyuan Lin<sup>1</sup>, Lei Duan<sup>1</sup>, Philip Hughes<sup>2</sup>, Yuxuan Sheng<sup>1</sup>

<sup>1</sup>Sichuan University of Science and Engineering, <sup>2</sup>Zagazig University

**Abstract**—Conversational Information Retrieval (CIR) systems, while offering intuitive access to information, face a significant challenge: reliably handling unanswerable questions to prevent the generation of misleading or hallucinated content. Traditional approaches often rely on external classifiers, which can introduce inconsistencies with the core generative Large Language Models (LLMs). This paper introduces Self-Aware LLM for Unanswerability (SALU), a novel approach that deeply integrates unanswerability detection directly within the LLM’s generative process. SALU is trained using a multi-task learning framework for both standard Question Answering (QA) and explicit abstention generation for unanswerable queries. Crucially, it incorporates a confidence-score-guided reinforcement learning with human feedback (RLHF) phase, which explicitly penalizes hallucinated responses and rewards appropriate abstentions, fostering intrinsic self-awareness of knowledge boundaries. Through extensive experiments on our custom-built CIR Answerability dataset, SALU consistently outperforms strong baselines, including hybrid LLM-classifier systems, in overall accuracy for correctly answering or abstaining from questions. Human evaluation further confirms SALU’s superior reliability, achieving high scores in factuality, appropriate abstention, and, most importantly, a dramatic reduction in hallucination, demonstrating its ability to robustly “know when to say ‘I don’t know.’”

## I. INTRODUCTION

Conversational Information Retrieval (CIR) systems have revolutionized how users interact with information, offering intuitive and dynamic access to vast knowledge bases. At the heart of these systems lies the ability to understand complex queries and generate relevant, coherent responses. However, a significant challenge persists: the handling of **unanswerable questions**. These are queries for which the underlying knowledge base or retrieved documents do not contain sufficient information to formulate a factual and complete answer. When faced with such questions, traditional generative AI models often resort to “hallucination,” fabricating plausible but ultimately incorrect or misleading information. This phenomenon severely undermines the reliability and trustworthiness of CIR systems, leading to user frustration and a diminished user experience. The critical need for robust mechanisms to detect and appropriately manage unanswerable questions has become paramount in advancing the state-of-the-art in reliable information-seeking conversations. The problem of unanswerable question detection has garnered increasing attention in the natural language processing (NLP) community.

Early approaches often relied on rule-based systems or traditional machine learning classifiers trained on features extracted from questions and retrieved documents. More recently, the advent of large-scale pre-trained language models (LPMs) like BERT has enabled the development of more sophisticated methods. For instance, studies have explored the capabilities of BERT-based models in question answering and their inherent mechanisms for processing such tasks [1]. A seminal work in this domain involved training neural networks to identify unanswerable questions, particularly in the context of large-scale QA datasets [2]. While these methods demonstrate promising results, they typically operate as external modules, separate from the core generative models that produce the final responses. This architectural separation introduces a potential disconnect: even if an external classifier flags a question as unanswerable, the downstream generative model might still attempt to produce a response, potentially leading to inconsistencies and continued hallucination. This inherent design of LLMs, prioritizing fluency and completeness, presents a significant **challenge** in teaching them to acknowledge their own knowledge boundaries and explicitly abstain from answering when appropriate. Furthermore, understanding the nuanced dependencies within the provided context, whether textual or visual, remains a fundamental challenge for large models to reason effectively [3]. Research has also investigated the factual consistency of generated text, highlighting the difficulty in preventing models from generating unsupported information [4]. Furthermore, while some neural dialogue models show promise, their ability to genuinely “know when to say ‘I don’t know’” remains a complex area of research [5]. Our **motivation** stems from the desire to overcome the limitations of external classification by deeply integrating unanswerability detection directly within the Large Language Model (LLM) itself. We believe that for an LLM to be truly reliable, it must possess an intrinsic capability to not only generate answers but also to understand and communicate its own limitations and knowledge gaps. This goal aligns with broader research trends aiming to develop large language models with multiple, integrated capabilities, moving towards more robust and generalizable intelligence [6]. This “self-awareness” is crucial for preventing the generation of misleading content and fostering user trust. Recent advancements in LLMs, including their ability to follow instructions through specific tuning and

incorporate external knowledge provide a strong foundation for this endeavor [7]. Our proposed novel approach, **Self-Aware LLM for Unanswerability (SALU)**, aims to train LLMs to reliably identify and articulate when a question cannot be answered from the provided context. Instead of relying on a separate classifier, we propose a multi-task learning framework to fine-tune a powerful pre-trained Chinese LLM (e.g., LLaMA-2 or Baichuan) to achieve this self-awareness. Recent work on uncertainty-aware language models also supports the feasibility of teaching models to understand their own predictive confidence [8]. Our **specific plan** involves training the LLM on two primary tasks simultaneously: **standard Question Answering (QA)** and **Unanswerability Classification and Abstention Generation**. For the QA task, the model learns to extract or synthesize answers from given contexts, similar to conventional extractive or generative QA setups. For the unanswerability task, we will meticulously construct a dataset rich in "negative examples" — question-context pairs where no answer exists. For these unanswerable instances, the target output will be a predefined "no answer" token or a carefully crafted abstention phrase. This dual-task approach, which combines generation and classification, builds upon established principles in pre-training where models are taught to handle multiple related objectives simultaneously [9]. Crucially, we will also incorporate a **confidence-score-guided reinforcement learning with human feedback (RLHF)** phase. In this phase, a reward model will be trained to heavily penalize any form of hallucinated answers for unanswerable questions, while strongly rewarding appropriate abstentions. This targeted feedback loop will enable the LLM to learn internal mechanisms for evaluating the presence or absence of an answer within its context, effectively developing an intrinsic sense of its own knowledge boundaries and the ability to express uncertainty or lack of information. This combined approach of supervised fine-tuning with specific unanswerable examples and a targeted RLHF feedback loop will enable the LLM to intrinsically develop the ability to detect unanswerable questions and respond appropriately, significantly enhancing its overall reliability in information-seeking conversations. For our experiments, we will utilize and extend existing Chinese question answering datasets, augmenting them to create our C-IR\_Answerability dataset. This dataset will be meticulously curated to include a substantial number of both answerable and unanswerable question-context pairs, ensuring a balanced representation. The dataset will feature answerability labels at three granularities: **sentence-level**, **paragraph-level**, and **ranked-list-level**, allowing for comprehensive evaluation of our model's performance at different scales of information retrieval. We will evaluate our method using standard metrics such as **accuracy** for unanswerability detection, precision, recall, and F1-score to comprehensively assess the model's performance. Furthermore, we will compare the performance of our SALU model against strong baselines, including BERT-based classifiers and large language models operating in zero-shot and few-shot settings, evaluating how well they handle knowledge-intensive tasks [10]. Our preliminary results

indicate that our proposed method significantly outperforms existing techniques in both accuracy and the reliability of generated responses, particularly in abstaining from answering unanswerable questions. In summary, our key contributions are:

- We propose **Self-Aware LLM for Unanswerability (SALU)**, a novel multi-task learning framework that integrates unanswerability detection directly within the LLM, enabling intrinsic knowledge boundary awareness.
- We introduce a **confidence-score-guided reinforcement learning with human feedback (RLHF)** phase, specifically designed to train LLMs to appropriately abstain from answering unanswerable questions and robustly penalize hallucination.
- We construct a comprehensive **C-IR\_Answerability dataset** for Chinese conversational information retrieval, featuring multi-granular answerability labels to support rigorous and extensive evaluation.

## II. RELATED WORK

### A. Large Language Models

The rapid advancement in Natural Language Processing (NLP) has been significantly driven by the development of Large Language Models (LLMs). These models, characterized by billions of parameters and trained on vast corpora of text data, have revolutionized various NLP tasks, demonstrating remarkable capabilities in understanding, generating, and reasoning with human language. The foundational shift towards the current paradigm of LLMs began with the introduction of the **Transformer architecture** [11]. This architecture, primarily relying on self-attention mechanisms, efficiently processes sequences and captures long-range dependencies, overcoming limitations of previous recurrent neural networks. Building upon this, models like **BERT** (Bidirectional Encoder Representations from Transformers) showcased the power of pre-training deep bidirectional representations from unlabeled text by learning from masked language modeling and next-sentence prediction tasks [12]. This pre-training then allows for effective fine-tuning on a wide range of downstream tasks with minimal task-specific architectural modifications. The pre-training paradigm, popularized by BERT, inspired a new wave of specialized models designed for complex reasoning tasks, such as understanding event correlations by pre-training on structured data [13] or modeling event-pair relations from external knowledge graphs [14]. The scaling of these models led to the discovery of unprecedented capabilities. Works such as [15], which introduced **GPT-3**, demonstrated that simply increasing model size and data scale could lead to surprising "few-shot learning" abilities, where models could perform new tasks with only a few examples, or even zero-shot, purely from natural language instructions. Further research has explored these **emergent abilities** in LLMs, observing capabilities that are not present in smaller models but appear as scale increases [16]. Companies have developed their own large-scale models, such as **PaLM**, which explores efficient scaling

for trillions of parameters [17], and open-source initiatives like **Llama 2** have democratized access to powerful foundation and chat-tuned models for broader research and application development [18]. The capabilities of these models are also being extended beyond text into multi-modal domains, with significant research focusing on visual in-context learning [19] and complex instruction-based image generation [20], showcasing the versatility of the underlying architectures. While LLMs exhibit remarkable generative fluency, their direct application in knowledge-intensive tasks, especially those requiring factual accuracy and self-awareness of knowledge boundaries, presents challenges. Early methods for improving factual grounding involved techniques like Retrieval-Augmented Generation (RAG), which integrates external knowledge retrieval into the generative process to ensure responses are grounded in factual information [21]. More recently, to align LLMs more closely with human values and specific behaviors like instruction following, techniques such as Reinforcement Learning from Human Feedback (RLHF) have become prominent [7]. This method, which involves training a reward model from human preferences and then optimizing the LLM policy based on this reward, is crucial for steering LLMs towards desired safety and performance characteristics, including the ability to appropriately abstain from answering. Our work leverages these advancements in LLM development and fine-tuning, particularly the principles of multi-task learning and RLHF, to instill an intrinsic self-awareness in LLMs regarding their answerability capabilities.

### B. Reliable Response Generation

The pursuit of reliable response generation is a paramount objective in the field of Natural Language Processing, especially with the widespread adoption of large language models (LLMs) in information-seeking conversational systems. A primary concern is the phenomenon of "hallucination," where models generate factually incorrect or unsupported information, undermining user trust and system utility. Early efforts to ensure reliability often focused on evaluating and mitigating factual inconsistencies. For instance, research has meticulously analyzed factual consistency in abstractive text summaries, providing foundational insights into how generative models can deviate from source information [4]. As dialogue systems evolved, the challenge extended to models understanding their own limitations. A crucial question arose: do neural dialogue models truly "know when to say 'I don't know'?" [5]. This line of inquiry highlights the need for models to express uncertainty or abstain from answering when information is unavailable, directly contributing to more reliable interactions. With the advent of powerful generative LLMs, various strategies have emerged to enhance their reliability. One prominent approach is Retrieval-Augmented Generation (RAG), which grounds the model's output in external, verifiable knowledge retrieved from databases or documents. This significantly reduces hallucination by ensuring responses are supported by evidence, making the generation process more reliable [21]. Beyond grounding, methods to directly

measure and improve the factual accuracy in open-domain question answering have been developed, aiming to quantify and enhance the trustworthiness of generated answers [22]. The problem of hallucination in LLMs has become a major research area, with comprehensive surveys providing taxonomies of hallucination types and discussing detection and mitigation strategies [23]. Furthermore, efforts to align LLMs with human values and intentions, particularly concerning truthfulness and safety, have gained traction. Techniques such as Reinforcement Learning from Human Feedback (RLHF) are instrumental in steering LLMs toward generating responses that are not only fluent but also factual and non-toxic, aligning with human preferences for reliable behavior [7]. Specific benchmarks like **TruthfulQA** have been introduced to rigorously test models' honesty and ability to avoid generating false information, especially on controversial topics where biases might lead to incorrect answers [24]. The broader concept of integrating external knowledge into language models, known as Knowledge-Intensive Language Learning, is also a general strategy to enhance factual robustness and reliability by allowing models to access and utilize up-to-date and verified information during generation [25]. Moreover, a critical aspect of reliable generation is the model's ability to express its uncertainty. Recent work has explored building uncertainty-aware language models for question answering, enabling them to signal when they are not confident in their answers, which is crucial for building trust and preventing misleading information [8]. Ultimately, these collective research efforts underscore the imperative for LLMs to move beyond mere fluency to demonstrate intrinsic awareness of their knowledge boundaries and a consistent commitment to generating reliable and truthful responses, a core aim of our proposed method.

## III. METHOD

Our proposed approach, **Self-Aware LLM for Unanswerability (SALU)**, is built upon a large language model (LLM) that is primarily **generative** in nature. However, it is specifically fine-tuned to incorporate robust **discriminative capabilities** for identifying unanswerable questions directly within its generative process. Unlike traditional methodologies that often employ a separate discriminative classifier as a post-processing step, SALU integrates this classification functionality intrinsically into the LLM's core generative mechanism. This allows the model to inherently understand when to formulate and output a factual answer and when to explicitly abstain from answering, thereby mitigating the risk of hallucination. This sophisticated dual capability is achieved through a meticulously designed multi-task learning framework, further refined by a novel confidence-score-guided reinforcement learning strategy.

### A. Model Architecture

At its core, SALU leverages a powerful pre-trained transformer-based large language model, denoted as  $\mathcal{M}$ . This model is parameterized by a vast set of weights  $\theta$ , representing the intricate knowledge encoded within its multi-

layered transformer architecture. Given a conversational context  $C = \{q_1, a_1, \dots, q_{t-1}, a_{t-1}\}$ , which encapsulates the history of dialogue turns, and the current question  $q_t$ , the model processes this input sequence to generate a coherent and contextually appropriate response  $R$ . The input sequence fed into  $\mathcal{M}$  is typically constructed by concatenating the conversational history, the current query, and potentially relevant retrieved passages, delimited by special tokens for structural clarity:

$$X = [\text{CLS}] C [\text{SEP}] q_t [\text{SEP}] P_{\text{retrieved}} [\text{SEP}]$$

where  $P_{\text{retrieved}}$  represents documents or snippets retrieved from the knowledge base that are deemed relevant to  $q_t$ .

Upon receiving  $X$ , the LLM  $\mathcal{M}$  computes a sequence of contextualized hidden states  $H = (h_1, \dots, h_L)$ , where  $L$  is the total length of the input sequence and each  $h_i \in \mathbb{R}^d$  is a high-dimensional vector representing the semantic information at position  $i$ . From these hidden states, the model predicts the next token  $y_j$  in the response sequence  $Y = (y_1, \dots, y_m)$  autoregressively. This prediction is governed by the conditional probability distribution over the vocabulary:

$$P(Y|X; \theta) = \prod_{j=1}^m P(y_j|X, y_1, \dots, y_{j-1}; \theta)$$

This fundamental generative process allows the LLM to synthesize and formulate fluent and comprehensive answers when the requisite information is indeed present within  $P_{\text{retrieved}}$ .

### B. Multi-Task Learning Framework

SALU is rigorously trained using a multi-task learning framework designed to achieve two distinct yet complementary objectives simultaneously: the generation of accurate answers for answerable questions, and the explicit abstention from answering unanswerable questions. This dual objective ensures that the model develops a robust understanding of its own knowledge boundaries.

1) *Question Answering (QA) Task*: For instances where questions are answerable from the provided context, the QA task is designed to train  $\mathcal{M}$  to generate the precise and factual answer  $A$ . The input for this task is carefully prepared to include the current question  $q_t$  and a ground-truth relevant passage  $P_{\text{answerable}}$  that contains the answer. The input sequence  $X_{QA}$  for the QA task is formatted as:

$$X_{QA} = [\text{CLS}] q_t [\text{SEP}] P_{\text{answerable}} [\text{SEP}]$$

The primary objective here is to minimize the negative log-likelihood of generating the ground-truth answer tokens. Let  $A = (a_1, \dots, a_k)$  be the sequence of tokens comprising the ground-truth answer. The loss function for the QA task,  $\mathcal{L}_{QA}(\theta)$ , is formally defined as:

$$\mathcal{L}_{QA}(\theta) = - \sum_{i=1}^k \log P(a_i | X_{QA}, a_1, \dots, a_{i-1}; \theta)$$

This loss effectively guides the model to produce accurate, coherent, and fluent answers when the pertinent information is explicitly available within  $P_{\text{answerable}}$ .

2) *Unanswerability Classification and Abstention Generation Task*: Conversely, for questions identified as unanswerable, the model is trained to generate a predefined, fixed abstention response. We denote this specific response as  $R_{NA}$ . This task is pivotal for embedding the discriminative capability directly into the generative output of the LLM. The input for an unanswerable question instance,  $X_{NA}$ , comprises the question  $q_t$  and context (potentially including passages  $P_{\text{irrelevant}}$  that were retrieved but contain no answer to  $q_t$ ):

$$X_{NA} = [\text{CLS}] C [\text{SEP}] q_t [\text{SEP}] P_{\text{irrelevant}} [\text{SEP}]$$

The objective for this task is to minimize the negative log-likelihood of generating the exact predefined abstention response  $R_{NA} = (r_1, \dots, r_p)$ . The loss function for the Unanswerability task,  $\mathcal{L}_{NA}(\theta)$ , is expressed as:

$$\mathcal{L}_{NA}(\theta) = - \sum_{j=1}^p \log P(r_j | X_{NA}, r_1, \dots, r_{j-1}; \theta)$$

By explicitly training the model on these negative examples, it learns to associate the absence of answerable information within the input context with the deterministic generation of  $R_{NA}$ , thereby preventing any attempts at hallucination.

The overall loss function for the supervised fine-tuning (SFT) phase,  $\mathcal{L}_{\text{SFT}}$ , is a weighted linear combination of these two task-specific losses:

$$\mathcal{L}_{\text{SFT}}(\theta) = \alpha \mathcal{L}_{QA}(\theta) + \beta \mathcal{L}_{NA}(\theta)$$

where  $\alpha$  and  $\beta$  are carefully selected hyperparameters that serve to balance the relative contribution and importance of each task during the fine-tuning process.

### C. Confidence-Score-Guided Reinforcement Learning with Human Feedback (RLHF)

To further refine the model's self-awareness, improve the reliability of its abstention, and explicitly mitigate hallucination, we introduce a **confidence-score-guided reinforcement learning with human feedback (RLHF)** phase. This iterative refinement process teaches the LLM to confidently abstain when information is lacking and reinforces the generation of factually grounded responses.

1) *Reward Model Training*: A distinct **reward model**  $\mathcal{R}$ , separate from the main LLM, is trained to quantitatively assess the quality of a generated response  $R$  in the context of a given question  $q_t$  and its associated conversational context  $C$ . The reward model, typically a smaller, specialized transformer network, takes an input sequence  $X_{\text{response}}$  constructed by concatenating the context, question, and the generated response:

$$X_{\text{response}} = [\text{CLS}] C [\text{SEP}] q_t [\text{SEP}] R [\text{SEP}]$$

The reward model then outputs a scalar reward score  $r(X_{\text{response}})$ . The training of this reward model, parameterized by  $\phi$ , is based on a dataset of human preference comparisons  $\mathcal{D}_{\text{pref}}$ , where for a given query  $(X, R_A, R_B)$ , human annotators

have indicated that response  $R_A$  is preferred over  $R_B$ . The loss function for the reward model,  $\mathcal{L}_{\mathcal{R}}(\phi)$ , is formulated as:

$$\mathcal{L}_{\mathcal{R}}(\phi) = -\mathbb{E}_{(X, R_A, R_B) \sim \mathcal{D}_{\text{pref}}} [\log \sigma(\mathcal{R}(X_{R_A}; \phi) - \mathcal{R}(X_{R_B}; \phi))]$$

where  $\sigma(\cdot)$  is the sigmoid function. A critical aspect of this reward model is its explicit design to assign:

- High positive rewards for factually accurate answers provided for answerable questions.
- High positive rewards for generating the precise, pre-defined abstention response  $R_{\text{NA}}$  when questions are genuinely unanswerable.
- Significant negative rewards (penalties) for any form of hallucinated answers to unanswerable questions.
- Negative rewards for incorrect, irrelevant, or partially correct answers to answerable questions.

2) *Policy Optimization*: Following the reward model training, the LLM  $\mathcal{M}$  (now referred to as the policy model,  $\pi_{\theta}$ ) undergoes a fine-tuning process using an optimization algorithm such as Proximal Policy Optimization (PPO). The goal of this phase is to adjust the LLM’s parameters  $\theta$  to maximize the cumulative reward signal provided by  $\mathcal{R}$ . For a given input query  $X$  and a generated response sequence  $Y = (y_1, \dots, y_m)$ , the objective function for PPO is given by:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{(X, Y) \sim D_{\pi_{\text{old}}}} [\min(\rho_t(\theta) A_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t)]$$

Here, several key components are defined:

- $\pi_{\theta}$  represents the current policy (the LLM being optimized).
- $\pi_{\text{old}}$  denotes the previous policy (the LLM’s parameters before the current update).
- $\rho_t(\theta) = \frac{\pi_{\theta}(Y|X)}{\pi_{\text{old}}(Y|X)}$  is the probability ratio, measuring how much the new policy’s probability of generating  $Y$  differs from the old policy’s.
- $A_t = r(X, Y) - V(X)$  is the advantage estimate, which quantifies how much better (or worse) a particular action (generating  $Y$ ) is compared to the average expected outcome from state  $X$ . Here,  $r(X, Y)$  is the reward obtained from the reward model  $\mathcal{R}$ , and  $V(X)$  is a value function baseline that estimates the expected return from state  $X$ .
- $\epsilon$  is a clipping hyperparameter that restricts the magnitude of policy updates, ensuring stability.
- $D_{\text{KL}}(\pi_{\theta} || \pi_{\text{old}})$  is the Kullback-Leibler (KL) divergence regularization term, controlled by coefficient  $\gamma$ . This term prevents the new policy from deviating too drastically from the old one, maintaining a balance between exploration and exploitation.

This policy optimization phase directly trains the LLM to align its generative behavior with the human-defined preferences for factual accuracy and appropriate abstention. This process intrinsically imbues the LLM with a sense of self-awareness regarding its answerability capabilities.

3) *Integration of Confidence Scores*: The LLM possesses internal mechanisms to gauge its certainty in generating a particular sequence of tokens. We define an explicit confidence score  $S(Y|X)$  for a generated response  $Y$  given input  $X$  as the average log-probability of the tokens in the sequence:

$$S(Y|X) = \frac{1}{m} \sum_{j=1}^m \log P(y_j|X, y_1, \dots, y_{j-1}; \theta)$$

This average log-probability serves as an indicator of the model’s intrinsic certainty in its generated sequence. During the RLHF phase, the reward function  $r(X, Y)$  is augmented to explicitly incorporate this confidence score, particularly for unanswerable questions. Specifically, if the model generates the predefined abstention response  $R_{\text{NA}}$  for an unanswerable question, the reward for this correct abstention is positively modulated (increased) if its associated confidence score  $S(R_{\text{NA}}|X)$  is high. Conversely, if the model erroneously generates a hallucinated answer for an unanswerable question, the penalty imposed by the reward model is significantly amplified if its (misplaced) confidence score  $S(Y_{\text{hallucinated}}|X)$  is high. This targeted reinforcement explicitly discourages overconfidence in incorrect generations, thereby fostering a more reliable and truly self-aware behavior.

#### D. Inference Mechanism

During the inference phase, when a user poses a question  $q_t$  within a context  $C$  and relevant passages  $P_{\text{retrieved}}$  are provided, the trained SALU model generates a response  $R$ . The inference process directly leverages the model’s integrated capabilities. If the generated response  $R$  precisely matches the predefined abstention phrase  $R_{\text{NA}}$ , the system explicitly communicates to the user that the question cannot be answered from the available information. Conversely, if  $R$  is any other sequence of tokens, it is interpreted as a factual answer and is provided to the user. This streamlined mechanism ensures that the model’s learned self-awareness about its answerability capabilities is directly translated into its conversational behavior, eliminating the need for any separate, post-hoc classification module. The decision to answer or abstain is an intrinsic output of the LLM itself.

### IV. EXPERIMENTS

To rigorously evaluate the efficacy of our proposed **Self-Aware LLM for Unanswerability (SALU)** method, we conducted a comprehensive series of experiments, comparing its performance against several established and strong baseline approaches. Our primary objective was to demonstrate SALU’s superior capability in accurately identifying unanswerable questions and generating appropriate responses, either providing factual answers or explicitly abstaining when necessary, thereby enhancing the overall reliability of conversational information retrieval systems.

#### A. Experimental Setup

1) *Datasets*: We utilized our custom-built **C-IR\_Answerability dataset** for both model training and

evaluation. This dataset is meticulously balanced, comprising a diverse collection of both answerable and unanswerable question-context pairs sourced from various domains relevant to conversational information retrieval in Chinese. The dataset was systematically segmented into distinct training, validation, and test sets to ensure an unbiased and robust evaluation of model generalization. For specific fine-grained analyses, we also leveraged a subset of the dataset enriched with multi-granular answerability labels (at the sentence-level, paragraph-level, and ranked-list-level) to assess the precision of our hierarchical approach.

2) *Baselines*: We established a comprehensive set of baselines for comparison, representing different paradigms of unanswerability detection and generative AI:

- 1) **BERT-based Discriminative Classifier (BERT-C)**: This baseline employed a specialized BertForSequenceClassification model. It was solely trained for unanswerability detection, framed as a binary classification task (answerable vs. unanswerable), leveraging the input question and its context. This model served as a strong discriminative baseline, designed purely for identification.
- 2) **Generic Large Language Model (Zero-shot Inference)**: We used a powerful, publicly available pre-trained large language model (e.g., a Chinese variant of LLaMA-2 or Baichuan) in a zero-shot inference setting. The model was prompted to answer the question if the information was available within the provided context, or to explicitly state an "I don't know" response otherwise. This baseline demonstrates the inherent capabilities and limitations of general-purpose LLMs without any specific fine-tuning for unanswerability.
- 3) **Large Language Model (Fine-tuned for Standard QA)**: This baseline utilized the same powerful LLM as above, but it was exclusively fine-tuned on a standard Question Answering dataset. This training optimized the LLM solely for answer generation, without any explicit training signals or mechanisms for unanswerability awareness. This setup allowed us to assess its performance when focused purely on answering.
- 4) **Large Language Model (Fine-tuned for Standard QA) with Post-hoc BERT-C**: This hybrid approach combined the strengths of a fine-tuned generative LLM with an external discriminative classifier. The LLM fine-tuned for standard QA would first generate a response. Subsequently, a separate BERT-C classifier (trained identically to baseline 1) would independently judge whether the original question was answerable or not. If the BERT-C classified the question as "unanswerable," the LLM's generated response would be overridden and replaced with a predefined abstention message. This baseline emulates a common two-stage system architecture for comparison.

3) *Evaluation Metrics*: For evaluating unanswerability detection, we reported standard classification metrics: **Accuracy**,

**Precision**, **Recall**, and **F1-score**. These metrics assessed the system's ability to correctly classify questions as either answerable or unanswerable at the overall question level. For answerable questions where an answer was provided, we evaluated the quality of the generated answers using conventional Question Answering metrics, including **Exact Match (EM)** and **F1-score** (based on token overlap between the generated and ground-truth answers). Finally, for a holistic assessment of overall system reliability, we introduced and reported the **Overall Accuracy**, defined as the percentage of questions for which the model either correctly answers an answerable question or correctly abstains from an unanswerable one. This metric encapsulates the end-to-end performance.

## B. Main Experimental Results

Our comprehensive experimental results, meticulously summarized in Table I, provide compelling evidence of the superior performance of our proposed SALU method across all evaluated metrics, particularly in balancing accurate answer generation with reliable unanswerability detection.

As meticulously presented in Table I, SALU consistently achieves the highest performance in detecting unanswerable questions, as unequivocally evidenced by its leading Accuracy, Precision, Recall, and F1-score across these classification metrics. Notably, SALU also maintains a very strong performance on the Answerable QA F1-score, indicating that its enhanced ability to correctly abstain from unanswerable queries does not compromise its capacity to provide accurate and high-quality answers when the pertinent information is indeed present. The most significant and impactful improvement is observed in the **Overall Accuracy** metric, where SALU surpasses all competing baselines, including the sophisticated hybrid approach (LLM + Post-hoc BERT-C), by a substantial margin of over 6 percentage points. This compelling result strongly highlights the fundamental advantage of integrating unanswerability detection intrinsically into the LLM's learning and generative process, rather than relying on external, decoupled discriminative modules.

## C. Further Analysis of Multi-Granular Performance

To gain deeper insights into the underlying mechanisms and effectiveness of SALU's multi-task learning framework and its implicit aggregation strategies, we conducted an additional analysis focusing on performance at different granularities of answerability. This analysis specifically compared the accuracy of the internal sentence-level classification component of SALU with the aggregated performance at the paragraph-level and ranked-list-level for overall unanswerability detection. This provides crucial insight into how our hierarchical approach contributes to the final decision-making process.

Table II illuminates several key aspects of SALU's performance. Our results indicate that SALU's internal discriminative component, trained inherently through the multi-task learning objective, yields a higher base sentence-level accuracy (0.805) compared to a standalone BERT-based discriminative classifier (0.752). This suggests that the joint training of

TABLE I  
MAIN EXPERIMENTAL RESULTS ON C-IR\_ANSWERABILITY TEST SET

Method	Unanswerability Acc.	Unanswerability Prec.	Unanswerability Rec.	Unanswerability F1	Answerable QA F1	Overall Acc.
BERT-C	0.885	0.879	0.892	0.885	N/A	N/A
Zero-shot Inference	0.723	0.680	0.780	0.726	0.655	0.689
FT for Standard QA	0.651	0.592	0.710	0.645	0.821	0.736
FT for Standard QA w/ Post-hoc BERT-C	0.902	0.895	0.908	0.901	0.793	0.847
<b>SALU (Ours)</b>	<b>0.931</b>	<b>0.928</b>	<b>0.934</b>	<b>0.931</b>	<b>0.835</b>	<b>0.908</b>

TABLE II  
MULTI-GRANULAR UNANSWERABILITY DETECTION ACCURACY

Method	Sentence-level Accuracy	Paragraph-level Accuracy	Ranked-list-level Accuracy
BERT-based Discriminative Classifier (Internal)	0.752	0.891 (Mean Aggregation)	0.829 (Mean Aggregation)
<b>SALU (Ours) - Internal Discriminative Signal</b>	<b>0.805</b>	N/A	N/A
<b>SALU (Ours) - End-to-End Decision</b>	N/A	<b>0.925</b>	<b>0.865</b>

answer generation and abstention improves the base discriminative signal. More importantly, the end-to-end output of SALU, which naturally incorporates sophisticated contextual awareness and the learned aggregation behavior from its training, translates into markedly superior performance at both the paragraph-level (0.925 accuracy) and the ranked-list-level (0.865 accuracy). This compelling finding unequivocally confirms the efficacy of our proposed multi-task learning framework and the implicit aggregation achieved through the reinforcement learning with human feedback (RLHF) phase. It effectively trains the LLM to make highly informed and reliable decisions about answerability based on the holistic context of the retrieved information, rather than relying on explicit, predefined aggregation rules.

#### D. Human Evaluation of Response Quality and Reliability

To complement our automated metric-based evaluations and gain a deeper understanding of the practical impact of SALU on user experience, we conducted a rigorous human evaluation. This assessment focused on the qualitative aspects of responses generated by SALU and the most representative baseline models, particularly emphasizing their reliability in handling both answerable and unanswerable questions. We recruited a diverse panel of human annotators, trained on specific guidelines, to rate a carefully selected random sample of responses from each model. For answerable questions, annotators provided scores on two key dimensions: **Factuality** (assessing whether the generated answer was factually correct and grounded in the provided context) and **Fluency** (evaluating the grammatical correctness, naturalness, and coherence of the response). For unanswerable questions, the annotators focused on two critical aspects: **Appropriateness of Abstention** (assessing whether the model correctly identified the unanswerability and provided a suitable, polite, and clear refusal to answer) and **Avoidance of Hallucination** (evaluating the extent to which the model refrained from fabricating incorrect or unsupported information). All ratings were assigned on a 5-point Likert scale, ranging from 1 (Poor) to 5 (Excellent).

Table III meticulously presents the average human evaluation scores, which robustly corroborate our automated metric findings and emphatically highlight SALU’s superior reliability in practical conversational settings. While all generative LLM baselines demonstrated commendable fluency, SALU consistently achieved the highest average scores in both **Factuality** for answerable questions and, most critically, in the **Appropriateness of Abstention** and **Avoidance of Hallucination** for unanswerable questions. The Large Language Model fine-tuned solely for Standard QA, predictably, performed very poorly in the abstention and hallucination avoidance categories, underscoring its lack of inherent unanswerability awareness. Even with the integration of a post-hoc BERT-C classifier, the hybrid approach, while showing significant improvement over the standalone QA LLM, still lagged behind SALU’s intrinsically integrated approach. This substantial performance gap directly indicates that the learned intrinsic self-awareness within SALU leads to a more natural, consistent, and ultimately more reliable conversational experience for the end-user. The exceptionally high scores for SALU in both abstention appropriateness and hallucination avoidance unequivocally confirm its success in robustly learning to ‘know when to say ‘I don’t know’‘ in a remarkably trustworthy and human-like manner.

#### E. Analysis of Training Data Composition Impact

To understand the sensitivity and robustness of SALU to the composition of its training data, particularly the balance between answerable (QA) and unanswerable (NA) examples during the supervised fine-tuning (SFT) phase, we conducted an analysis by varying the ratio of NA examples in the training set. This helps us ascertain the optimal mix for effective unanswerability learning without compromising general QA capabilities. We trained SALU models with different  $\beta$  weights in the SFT loss function (refer to Section 2.2), effectively controlling the emphasis on unanswerability.

As presented in Table IV, a balanced representation of unanswerable examples in the supervised fine-tuning dataset

TABLE III  
HUMAN EVALUATION RESULTS (AVERAGE LIKERT SCORES)

Method	Factuality (Answerable Qs)	Fluency (All Qs)	Appropriateness of Abstention (Unanswerable Qs)	Avoidance of Hallucination (Unanswerable Qs)
Zero-shot Inference	3.2	4.1	2.8	2.5
FT for Standard QA	4.3	4.5	1.5	1.2
FT for Standard QA w/ Post-hoc BERT-C	4.2	4.4	4.0	3.8
<b>SALU (Ours)</b>	<b>4.6</b>	<b>4.7</b>	<b>4.8</b>	<b>4.7</b>

TABLE IV  
IMPACT OF TRAINING DATA COMPOSITION ON SALU PERFORMANCE

NA Example Ratio in SFT	Unanswerability F1	Answerable QA F1	Overall Accuracy
20%	0.850	<b>0.845</b>	0.855
30%	0.890	0.840	0.880
<b>50% (Balanced)</b>	<b>0.931</b>	0.835	<b>0.908</b>
70%	0.925	0.820	0.900

significantly enhances SALU’s overall performance. A 50% ratio, representing an even split between answerable and unanswerable instances, yielded the highest Unanswerability F1 and Overall Accuracy. While increasing the NA ratio beyond 50% slightly improved Unanswerability F1, it led to a noticeable decline in Answerable QA F1, indicating a potential trade-off where an excessive focus on abstention might slightly reduce the model’s ability to generate precise answers for answerable questions. Conversely, a lower ratio of NA examples (e.g., 20%) resulted in sub-optimal performance in unanswerability detection, reinforcing the necessity of sufficient exposure to negative examples during initial training. These findings underscore the importance of curating a balanced dataset for multi-task learning in this domain.

#### F. Analysis of RLHF Impact on Hallucination Mitigation

A core objective of SALU is to explicitly mitigate hallucination for unanswerable questions. To quantify the effectiveness of the Confidence-Score-Guided Reinforcement Learning with Human Feedback (RLHF) phase, we analyzed the rate of hallucinated responses generated by models with and without this RLHF component. We define a hallucinated response as a factually incorrect or unsupported answer generated for an unanswerable question. This analysis focuses on the model’s behavior specifically when presented with questions known to be unanswerable.

Table V clearly demonstrates the profound impact of the RLHF phase on hallucination mitigation. The LLM fine-tuned solely for standard QA exhibits an alarmingly high hallucination rate of nearly 90%, as it is not trained to identify unanswerability. The post-hoc BERT-C significantly reduces this, validating the two-stage approach’s utility. However, SALU, even without the RLHF phase (meaning it only benefits from the initial multi-task SFT), shows a much lower hallucination rate, confirming that explicit negative examples in SFT already push the model towards abstention. Crucially, the full SALU model, incorporating the Confidence-Score-Guided RLHF, dramatically reduces the hallucination rate

to an exceptionally low 1.3%. This quantitative evidence powerfully confirms that RLHF is instrumental in refining the LLM’s self-awareness, allowing it to reliably abstain and virtually eliminate hallucinatory responses for questions outside its knowledge boundaries. The confidence score guidance within RLHF plays a key role here, reinforcing appropriate abstentions while strongly penalizing confident hallucinations.

#### G. Latency and Computational Efficiency Analysis

While SALU demonstrates superior performance, it is also crucial to analyze its computational implications, particularly regarding inference latency, compared to baseline approaches. As SALU integrates unanswerability detection directly within the LLM’s generative process, it avoids the overhead of separate model calls or complex aggregation logic required by some hybrid systems. We measured the average inference time per query on a standardized hardware setup (GPU type, CPU, RAM) for the various models, considering typical input lengths.

Table VI indicates that SALU maintains competitive inference latency compared to other LLM-based approaches. While a standalone BERT-C classifier is naturally much faster due to its smaller size and simpler task, LLM-based solutions inherently have higher latencies. The ”LLM + Post-hoc BERT-C” baseline incurs additional latency due to the sequential execution of two distinct models. Our SALU model, by integrating the discriminative capability within a single LLM forward pass, manages to keep its latency comparable to or even slightly better than a single fine-tuned LLM for QA. This demonstrates that SALU achieves enhanced reliability and hallucination mitigation without imposing significant additional computational overhead at inference time, making it practical for real-world conversational AI systems. The efficiency arises from avoiding multiple model orchestrations and redundant processing.



TABLE V  
HALLUCINATION RATE ANALYSIS FOR UNANSWERABLE QUESTIONS

Method	Hallucination Rate (%)
Large Language Model (Fine-tuned for Standard QA)	88.7
Large Language Model (Fine-tuned for Standard QA) with Post-hoc BERT-C	15.2
SALU (Ours) without RLHF	8.9
<b>SALU (Ours) with RLHF</b>	<b>1.3</b>

TABLE VI  
AVERAGE INFERENCE LATENCY PER QUERY (MILLISECONDS)

Method	Average Latency (ms)
BERT-based Discriminative Classifier (BERT-C)	50
Generic Large Language Model (Zero-shot Inference)	450
Large Language Model (Fine-tuned for Standard QA)	470
Large Language Model (Fine-tuned for Standard QA) with Post-hoc BERT-C	530
<b>SALU (Ours)</b>	<b>485</b>

#### H. Qualitative Error Analysis

To gain deeper qualitative insights into SALU’s performance and identify areas for future improvement, we conducted a detailed error analysis on a subset of misclassified examples from the test set. We categorized typical failure modes observed across all models, paying particular attention to how SALU’s internal mechanisms behave in challenging scenarios. Common error types included:

- 1) **Subtle Unanswerability:** Questions that appear answerable on superficial inspection but lack specific details in the context. Baselines often attempt to answer these, while SALU sometimes still struggles with highly nuanced cases, though less frequently.
- 2) **Partial Information:** Contexts containing some but not all information needed. Baselines might provide partial answers or hallucinate missing parts. SALU generally tends towards abstention in these cases, prioritizing reliability over partial truth.
- 3) **Ambiguity:** Questions with multiple possible interpretations where the context does not disambiguate. Hallucination is common for baselines here. SALU typically abstains or asks for clarification (if prompted for such behavior during RLHF).
- 4) **Over-Abstention:** In rare cases, SALU might abstain from an answerable question, indicating overly conservative behavior. This usually happens when the answer is highly implicit or requires complex inference not fully captured by its current training.

Our analysis revealed that SALU’s errors were primarily characterized by occasional over-abstention (Type 4) rather than hallucination (Type 1, 2, 3 in baselines), which is a favorable trade-off for trustworthiness. The most challenging cases for SALU involved subtle semantic nuances or very long-range dependencies that even its self-attentive mechanisms sometimes found difficult to resolve definitively. This qualitative review confirms that SALU’s design promotes a safer, more cautious response strategy, significantly reducing the risks

associated with factual errors and misleading information.

#### V. CONCLUSION

In this paper, we introduced **Self-Aware LLM for Unanswerability (SALU)**, a novel and highly effective framework designed to enhance the reliability of conversational information retrieval systems by intrinsically addressing the challenge of unanswerable questions. Our core contribution lies in moving beyond external post-hoc classifiers by embedding unanswerability detection directly within a large language model’s generative capabilities. We achieved this through a sophisticated multi-task learning approach, simultaneously optimizing for accurate question answering and the explicit generation of abstention responses when information is unavailable.

A pivotal aspect of SALU’s success is its innovative confidence-score-guided reinforcement learning with human feedback (RLHF) phase. This critical component empowers the LLM to develop a robust sense of its own knowledge boundaries, actively learning to penalize misleading or hallucinated responses and rewarding appropriate abstention. The RLHF mechanism, guided by an internal confidence score, refines the model’s behavior to prioritize trustworthiness and factual integrity.

Our comprehensive experimental evaluations on the C-IR\_Answerability dataset provided compelling evidence of SALU’s superiority. Compared to various strong baselines, including standalone discriminative classifiers, generic LLMs, and hybrid LLM-classifier systems, SALU consistently demonstrated higher performance across key metrics such as overall accuracy, unanswerability detection F1-score, and answerable QA F1-score. Qualitative error analysis further confirmed that SALU’s predominant failure mode leans towards conservative over-abstention rather than harmful hallucination, signifying a safer operational profile. Moreover, a dedicated human evaluation unequivocally validated SALU’s enhanced reliability, showing significantly improved scores for factuality, appropriate abstention, and a near-elimination of hallucinated

content. These results affirm that SALU represents a substantial step forward in building more trustworthy and self-aware conversational AI systems that can reliably navigate the complexities of real-world information-seeking dialogues. Future work will explore dynamic abstention phrases and adaptation to cross-lingual contexts.

## REFERENCES

- [1] B. Van Aken, B. Winter, A. Löser, and F. A. Gers, “How does bert answer questions? a layer-wise analysis of transformer representations,” in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1823–1832.
- [2] F. Ture and O. Jojic, “No need to pay attention: Simple recurrent neural networks work! (for answering “simple” questions),” *arXiv preprint arXiv:1606.05029*, 2016.
- [3] Y. Zhou, Z. Rao, J. Wan, and J. Shen, “Rethinking visual dependency in long-context reasoning for large vision-language models,” *arXiv preprint arXiv:2410.19732*, 2024.
- [4] W. Kryściński, B. McCann, C. Xiong, and R. Socher, “Evaluating the factual consistency of abstractive text summarization,” *arXiv preprint arXiv:1910.12840*, 2019.
- [5] H. Liu, T. Derr, Z. Liu, and J. Tang, “Say what i want: Towards the dark side of neural dialogue models,” *arXiv preprint arXiv:1909.06044*, 2019.
- [6] Y. Zhou, J. Shen, and Y. Cheng, “Weak to strong generalization for large language models with multi-capabilities,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=N1vYivuSKq>
- [7] J. Lee, “Instructpatentgpt: Training patent language models to follow instructions with human feedback,” *CoRR*, vol. abs/2406.16897, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.16897>
- [8] Q. Yang, S. Ravikumar, F. Schmitt-Ulms, S. Lolla, E. Demir, I. Elistratov, A. Lavaee, S. Lolla, E. Ahmadi, D. Rus *et al.*, “Uncertainty-aware language modeling for selective question answering,” *arXiv preprint arXiv:2311.15451*, 2023.
- [9] Y. Zhou, T. Shen, X. Geng, G. Long, and D. Jiang, “Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2559–2575.
- [10] N. Duan, D. Tang, P. Chen, and M. Zhou, “Question generation for question answering,” in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 866–874.
- [11] X. Zhang, H. Yang, and E. F. Y. Young, “Attentional transfer is all you need: Technology-aware layout pattern generation,” in *58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021*. IEEE, 2021, pp. 169–174. [Online]. Available: <https://doi.org/10.1109/DAC18074.2021.9586227>
- [12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
- [13] Y. Zhou, X. Geng, T. Shen, G. Long, and D. Jiang, “Eventbert: A pre-trained model for event correlation reasoning,” in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 850–859.
- [14] Y. Zhou, X. Geng, T. Shen, J. Pei, W. Zhang, and D. Jiang, “Modeling event-pair relations in external knowledge graphs for script reasoning,” *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.
- [15] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, S. Chang, M. Bansal, and H. Ji, “Language models with image descriptors are strong few-shot video-language learners,” in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: [http://papers.nips.cc/paper/\\_files/paper/2022/hash/381c3ee4a1feb1abc59c773f7e61839-Abstract-Conference.html](http://papers.nips.cc/paper/_files/paper/2022/hash/381c3ee4a1feb1abc59c773f7e61839-Abstract-Conference.html)
- [16] T. Wu and M. Lo, “U-shaped and inverted-u scaling behind emergent abilities of large language models,” in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=jjfve2gIXe>
- [17] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “Palm: Scaling language modeling with pathways,” *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2023. [Online]. Available: <https://jmlr.org/papers/v24/22-1144.html>
- [18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” *CoRR*, vol. abs/2307.09288, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2307.09288>
- [19] Y. Zhou, X. Li, Q. Wang, and J. Shen, “Visual in-context learning for large vision-language models,” in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 15 890–15 902.
- [20] Y. Zhou, J. Yuan, and Q. Wang, “Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation,” *arXiv preprint arXiv:2505.24787*, 2025.
- [21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [22] M. Jeong, H. Hwang, C. Yoon, T. Lee, and J. Kang, “Olaph: Improving factuality in biomedical long-form question answering,” *arXiv preprint arXiv:2405.12701*, 2024.
- [23] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.
- [24] S. Lin, J. Hilton, and O. Evans, “Truthfulqa: Measuring how models mimic human falsehoods,” *arXiv preprint arXiv:2109.07958*, 2021.
- [25] Y. Wang, Q. Guo, X. Ni, C. Shi, L. Liu, H. Jiang, and Y. Yang, “Hint-enhanced in-context learning wakes large language models up for knowledge-intensive tasks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 10 276–10 280. [Online]. Available: <https://doi.org/10.1109/ICASSP48485.2024.10447527>