Attention-guided Self-reflection for Zero-shot Hallucination Detection in Large Language Models

Anonymous ACL submission

Abstract

Hallucination has emerged as a significant barrier to the effective application of Large Language Models (LLMs). In this work, we introduce a novel Attention-Guided SElf-Reflection (AGSER) approach for zero-shot hallucination detection in LLMs. The AGSER method utilizes attention contributions to categorize the input query into attentive and non-attentive queries. Each query is then processed separately through the LLMs, allowing us to compute consistency scores between the generated responses and the original answer. The difference between the two consistency scores serves as a hallucination estimator. In addition to its efficacy in detecting hallucinations, AGSER notably reduces computational overhead, requiring only three passes through the LLM and utilizing two sets of tokens. We have conducted extensive experiments with four widely-used LLMs across three different hallucination benchmarks, demonstrating that our approach significantly outperforms existing methods in zero-shot hallucination detection.

1 Introduction

Recently, Large Language Models (LLMs) (Zhao et al., 2023) have demonstrated superior ability and achieved excellent results in various natural language processing tasks, such as summarization (Ravaut et al., 2024), machine translation (Zhang et al., 2023a), autonomous agents (Wang et al., 2024), information retrieval (Xu et al., 2024), and knowledge graph reasoning (Sun et al., 2024). Despite the convenience offered by LLMs, they may produce overly confident answers that deviate from factual reality (Manakul et al., 2023; Zhang et al., 2023b; He et al., 2024). This is usually called the *Hallucination* phenomenon, which makes LLMs very untrustworthy (Zhang et al., 2023c; Li et al., 2024). This strongly limits the application of LLMs, especially in medical, financial, legal, and other scenarios. Thus, it is urgent to investigate the accurate and efficient hallucination detection in LLMs, and teach LLMs to say "I don't know" when they are not sure about the answers.

The most common hallucination detection methods are based on answer consistency (Manakul et al., 2023; Zhang et al., 2023b; Chen et al., 2024), in which the answers to the same query are sampled multiple times. Though effective, such methods heavily increase computation cost through multiple LLM running. They also rely on randomness, and when the LLM is extremely confident in the wrong answer, the same answer may be constantly generated during resampling (Zhang et al., 2023b). Moreover, none of the existing consistency-based approaches guides LLMs to rethink the answer generation process like humans do, which may help us to obtain a better consistency evaluation. Recently, more hallucination detection approaches have been proposed from other perspectives, but they require tool usage (Cheng et al., 2024), or annotated hallucination datasets (Azaria and Mitchell, 2023; He et al., 2024; Chuang et al., 2024a).

Considering that attention contributions in LLMs reflect the key parts of the answer generation process and provide hints about hallucinations (Yuksekgonul et al., 2024), we propose an Attention-Guided SElf-Reflection (**AGSER**) approach for zero-shot hallucination detection in LLMs, which refers to identifying hallucinations without requiring specific training on annotated samples from the target LLM. Specifically, according to attention contributions of tokens, we split the input query for LLMs into attentive and non-attentive queries. As the attentive query contains the major information for LLMs to generate the answer, if we input the attentive query into LLMs, the generated answer should be very similar to the original answer for a non-hallucination sample. On the other hand, due to language differences between attentive and original queries, the randomness of generating the hallucination answer has been enlarged, and we have a greater chance of detecting hallucination based on the inconsistency of answers. This is similar as when a human is doing reading comprehension, if asked to rethink about the answer, he or she will re-examine the attentive parts of the article, and may provide a new answer. Meanwhile, for a non-hallucination sample, there is almost no important information in the non-attentive query, and thus when we input the non-attentive query into LLMs, the generated answer should be extremely random and totally different from the original answer. In Sec. 4, we provide some observations to verify the above analysis.

Accordingly, in AGSER, we use attentive and non-attentive queries to guide LLMs to conduct self-reflection for hallucination detection. Specifically, we separately feed attentive and non-attentive queries into LLMs, and respectively calculate the consistency scores between the generated answers and the original answer, which are denoted as attentive and non-attentive consistency scores. Then, as smaller attentive consistency scores and larger non-attentive consistency scores indicate higher degrees of hallucination, we compute their difference as the hallucination estimator. This enables us to detect hallucinations in a zero-shot manner. Meanwhile, compared to conventional consistency-based approaches, AGSER reduces the computational overhead of resampling. It only requires three times of LLM running, and two times of token usage. We have conducted extensive experiments with four popular LLMs, and ASGER achieves state-of-the-art hallucination detection performances.

The main contributions of this work are summarized as follows:

- According to attention contributions of tokens in LLMs, we define attentive and non-attentive queries. For a hallucination sample, the generated answer of the attentive query has a larger chance to be different from the original answer, and the generated answer of the non-attentive query has a larger chance to be similar to the original answer.

- We propose a novel AGSER approach for zero-shot hallucination detection. AGSER uses attentive and non-attentive queries for constructing an effective hallucination estimator. It can also reduce the computational overhead of answer resampling.

- We have conducted extensive experiments with four popular LLMs, which demonstrate the effectiveness of our proposed AGSER approach in hallucination detection.

2 Related Work

Hallucination has become the major obstacle in constructing trustworthy LLMs (Zhang et al., 2023c). LLMs may generate overly confident non-factual contents. This brings great demand for automatic hallucination detection in LLMs (Li et al., 2024), especially in a zero-shot manner.

The most common hallucination detection approach is based on the inconsistency of the generated contents. SelfCheckGPT (Manakul et al., 2023) stochastically generates multiple responses besides the original answer, and detects the hallucination via verifying whether the responses support the original answer. SAC^3 (Zhang et al., 2023b) detects hallucinations through consistency analysis across different LLMs or cross rephrased queries. It also points out that generated answers to the same query may be consistent but non-factual. LogicCheckGPT (Wu et al., 2024) asks LLMs with questions with logical relationships for hallucination detection. INSIDE (Chen et al., 2024) attempts to calculate answer inconsistency in the sentence embedding space. InterrogateLLM (Yehuda et al., 2024) detects hallucinations via asking the reverse question, and verify whether the original question can be generated. Graph structure has also been extracted and applied for better estimation of answer consistency (Fang et al., 2025).

Moreover, the inner states of LLMs can tell hallucinations to some extent (Azaria and Mitchell, 2023). We can use hidden states (He et al., 2024) or attention values (Chuang et al., 2024a) for training classifiers to detect hallucinations. However, such approaches require training datasets, and may have trouble generalizing among different LLMs and different data (Orgad et al., 2024). Meanwhile, some works propose to call tools for constructing hallucination detectors (Cheng et al., 2024; Yin et al., 2023). In addition, some works attempt to refine LLM parameters to enhance the factuality, via aligning with factuality analysis results (Zhang et al., 2024b), truthful space editing (Zhang et al., 2024a), over-trust penalty (Leng et al., 2024), and confidence calibration (Liu et al., 2024). Contrastive decoding (Li et al., 2023; Chuang et al., 2024b; Leng et al., 2024; Cheng et al., 2025; Huo et al., 2025), which proposes to subtract output

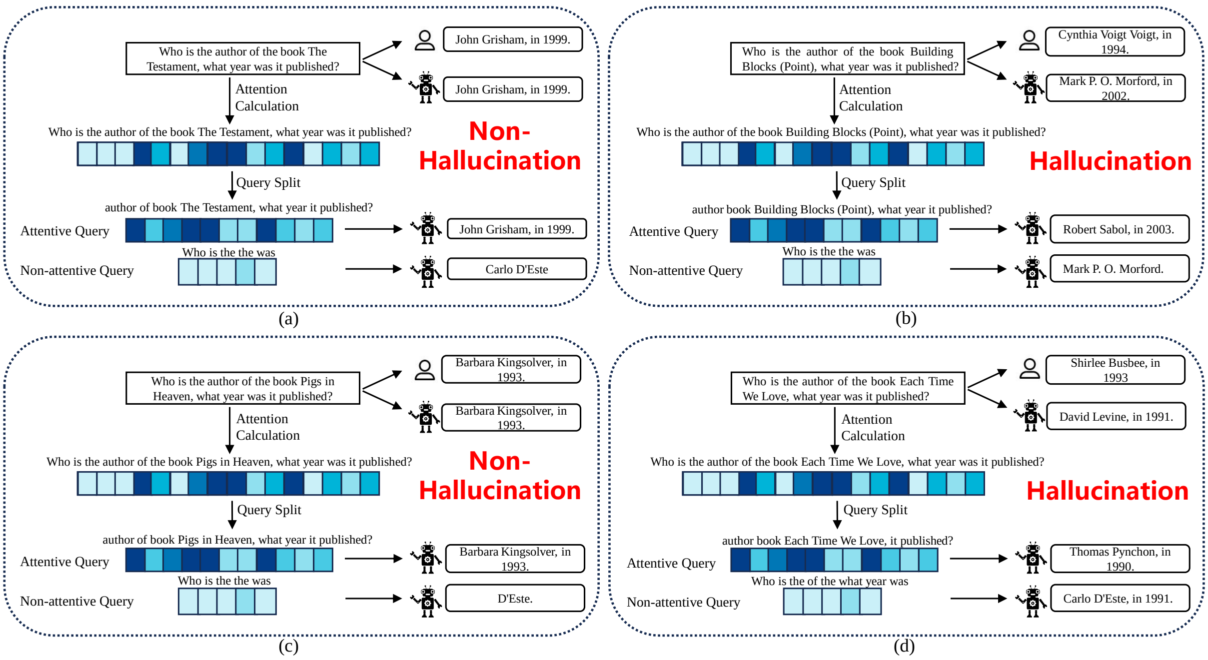


Figure 1: Some examples on feeding attentive and non-attentive queries into Llama2-7b. For non-hallucination samples, compared to the original answers, the answers of the attentive queries stay consistent, and those of the non-attentive queries otherwise. For hallucination samples, the answers of the attentive queries mostly change, and those of the non-attentive queries may remain unchanged.

logits with less factuality, has also been used for improving the factuality.

There is research showing that, LLMs' attention to some constraint tokens (such as important entities) relates to the factuality of the generated responses (Yuksekgonul et al., 2024). Accordingly, attention contributions can reflect the answer generation process of LLMs, and guide LLMs to conduct self-reflection for accurate hallucination detection.

3 Preliminary

A query is denoted as a sequence of tokens $X = \{x_1, x_2, ..., x_M\}$, in which x_i denotes the i-th token. We denote a LLM as $f(\bullet)$, and the generated answer is $Y = f(X)$. Specifically, the answer is a sequence of tokens $Y = \{y_1, y_2, ..., y_N\}$, in which y_j denotes the j-th token. Due to the hallucination phenomenon, Y may be factual or non-factual.

The self-attention layers are the core components in LLMs (Vaswani et al., 2017), and can reflect the key parts of the answer generation process of LLMs. We assume that the LLM has L self-attention layers and H heads. In the self-attention layers, there are two projection matrices $W_Q^{l,h}$ and $W_K^{l,h}$ for attention calculation, which denote query and key projections respectively, for layer l and head h, and the dimensionality $d_h = d/H$. The attention value matrix for layer l and head h can be calculated as

$$A^{l,h} = \sigma\left(\frac{\left(X^{l-1}W_Q^{l,h}\right)\left(X^{l-1}W_K^{l,h}\right)^\top}{\sqrt{d_h}}\right), \quad (1)$$

where σ denotes softmax function. And the attention contribution from token j to token i for layer l through all heads can be calculated as

$$a_{i,j}^l = \sum_{h=1}^{H} A_{i,j}^{l,h}. \quad (2)$$

Then, to obtain a score for measuring the contribution of the token i during the answer generation process of the LLM, we use the attention contribution from token i to the last token of the query as the token contribution score

$$s_i^l = a_{M,i}^l \,. \quad (3)$$

4 Analysis

To verify that we can use attention to guide LLMs to conduct self-reflection and accurately detect hallucinations, we present the following analysis. We adopt the attention at the middle layer, i.e., layer $L/2$, for the token contribution calculation. The contribution score at the middle layer

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.025	0.167	0.218	0.590
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.752	0.121	0.095	0.032

Table 1: Distribution of attentive consistency scores r^{att} with Llama2-7b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.845	0.121	0.031	0.003

Table 2: Distribution of non-attentive consistency scores r^{non_att} with Llama2-7b on the Books dataset.

for token i is $s_i^{mid} = a_{M,i}^{L/2}$, and the contribution scores for the entire input query are $S^{mid} = \{s_1^{mid}, ..., s_M^{mid}\}$. Then, we can split the input query $X = \{x_1, x_2, ..., x_M\}$ into attentive and non-attentive queries

$$X^{att} = \{x_i \,|\, s_i \in top_k\left(S\right)\}, \quad (4)$$

$$X^{non\_att} = \{x_i \,|\, s_i \notin top_k\left(S\right)\}, \quad (5)$$

where $s_i = s_i^{mid}$, $S = S^{mid}$, and $top_k\left(\bullet\right)$ means selecting tokens with k highest contributions. Here, we select the top $k = 2/3$ tokens. Then, we can obtain the corresponding responses of the LLM as $Y^{att} = f\left(X^{att}\right)$ and $Y^{non\_att} = f\left(X^{non\_att}\right)$. To measure the consistency between the attention-guided generated answers Y^{att}, Y^{non_att} and the original answer Y, we adopt the Rouge-L (Lin, 2004) similarity estimation [1], which provides an accurate evaluation for consistent answer pairs. Specifically, we have attentive consistency score and non-attentive consistency score as follows

$$r^{att} = Rouge\left(Y^{att}, Y\right), \quad (6)$$

$$r^{non\_att} = Rouge\left(Y^{non\_att}, Y\right). \quad (7)$$

To analyze the relationship between hallucinations in LLMs and attentive/non-attentive consistency scores, we conduct some pilot study on the Books dataset (Yehuda et al., 2024). We present the results with the Llama2-7b model (Touvron et al., 2023), which is a widely-used LLM. In Fig. 1, we illustrate four examples on feeding attentive and non-attentive queries into Llama2-7b. From the two non-hallucination samples we can observe that, the answers of the attentive queries stay consistent with the original answers, and the answers of the non-attentive queries are inconsistent with the original answers. Meanwhile, as shown in the two hallucination samples, the answers of the attention queries mostly change, while the answers of the non-attentive queries may remain unchanged. Furthermore, we show the distribution of attentive and non-attentive consistency scores in Tabs. 1 and 2 respectively. Obviously, the attentive consistency scores are much larger with non-hallucination samples than with hallucination samples. Specifically, most attentive consistency scores of non-hallucination samples are in $[0.75, 1.0]$, while most attentive consistency scores of hallucination samples are in $[0.0, 0.25)$. Moreover, non-attentive consistency scores of non-hallucination samples are all in $[0.0, 0.25)$, while hallucination samples have the chance to have larger non-attentive consistency scores. More results with other LLMs and on other datasets can be found in App. B. We can conclude that, smaller attentive consistency scores and larger non-attentive consistency scores indicate greater probabilities of hallucinations.

5 Methodology

According to the above analysis and conclusion, in this section, we introduce the AGSER approach for zero-shot hallucination detection in LLMs. The whole procedure is illustrated in Alg. 1.

In addition to adopting attention at the middle layer of a LLM for token contribution calculation as in Sec. 4, we can define the following token contribution scores

- **The first layer value**: $s_i^{first} = a_{M,i}^1$.
- **The middle layer value**: $s_i^{mid} = a_{M,i}^{L/2}$.
- **The last layer value**: $s_i^{last} = a_{M,i}^L$.
- **The maximum value of all layers**: $s_i^{max} = MAX\left(a_{M,i}^l | 0 < l \leq L\right)$.
- **The mean value of all layers**: $s_i^{mean} = MEAN\left(a_{M,i}^l | 0 < l \leq L\right)$.

[1] https://github.com/google-research/google-research/tree/master/rouge

Algorithm 1 The AGSER approach.

Input: A LLM $f(\bullet)$, and input query X.
Output: The degree of hallucination r.

1: Feed the query X into the LLM and obtain the answer $Y = f(X)$.
2: Calculate the attention contributions in the LLM as in Eq. 2, and obtain the token contribution scores $S = \{s_1, ..., s_M\}$.
3: According to S, select the top k tokens to construct the attentive query X^{att}, and the rest to form the non-attentive query X^{non_att} as in Eqs. 4 and 5.
4: Generate new answers $Y^{att} = f\left(X^{att}\right)$ and $Y^{non\_att} = f\left(X^{non\_att}\right)$.
5: Calculate attentive and non-attentive consistency scores r^{att} and r^{non_att} based on Rouge-L similarity estimation as in Eqs. 6 and 7.
6: Calculate the overall estimation of hallucination r as in Eq. 8.
7: **return** r.

Then, we can replace the token contribution score s_i in Eqs. 4 and 5 with the above scores for calculating the corresponding attentive and non-attentive queries X^{att} and X^{non_att}. And we can further obtain the attentive and non-attentive consistency scores r^{att} and r^{non_att} for estimating the degrees of hallucinations in LLMs as in Eqs. 6 and 7.

As smaller attentive consistency scores and larger non-attentive consistency scores indicate greater probabilities of hallucinations, we define the following score function as the final estimation of hallucinations in LLMs

$$r = \lambda r^{att} - r^{non\_att} \tag{8}$$

where λ denotes a hyper-parameter for balancing the attentive and non-attentive consistency scores. To be noted, with smaller scores, hallucinations are more severe, and LLMs may generate non-factual contents.

6 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of AGSER in zero-shot hallucination detection in LLMs.

6.1 Experimental Settings

Following (Yehuda et al., 2024), we conduct experiments on the **Books**, **Movies** and Global Country Information (**GCI**) datasets, which cover various domains. For the evaluation of hallucination detection results, the detection predictions are compared against the correctness of LLMs' answers. The correctness is determined as in (Yehuda et al., 2024) for samples from different datasets. More details of the datasets can be found in App. A. Meanwhile, we use the Area Under Curve (**AUC**) as the evaluation metric.

We compare the proposed AGSER approach with SBERT (Reimers and Gurevych, 2019), SelfCheckGPT (Manakul et al., 2023), INSIDE (Chen et al., 2024) and InterrogateLLM (Yehuda et al., 2024) in zero-shot hallucination detection. Introduction of these baselins can be found in App. C.

Moreover, we implement AGSER and other compared hallucination detection approaches with four popular and outstanding open-source LLMs: **Llama2-7b** [2], **Llama2-13b** [3] (Touvron et al., 2023), **Llama3-8b** [4] (Dubey et al., 2024), and **Qwen2.5-14b** [5] (Qwen, 2024). More details of these LLMs can be found in App. D.

For InterrogateLLM, we adopt the best version reported in the original paper, i.e., an ensemble of GPT-3 (Brown et al., 2020), Llama2-7b and Llama2-13b. For SelfCheckGPT, INSIDE and InterrogateLLM, we perform resampling of answers for 5 times to calculate the consistency scores.

In our proposed AGSER approach, we set $k = 2/3$ and $\lambda = 1.0$. And we adopt the mean value of all layers in a LLM, i.e., s_i^{mean}, for token contribution estimation. We have not tuned the hyper-parameters for the optimal results on each dataset for each LLM, cause it is usually impractical to obtain sufficient high-quality hallucination and non-hallucination samples specific to each LLM as validation samples. According to results in Sec. 6.4, with the above selected hyper-parameters, we can not achieve the optimal results, but the overall satisfactory results. Meanwhile, the prompts used in our experiments are illustrated in App. F.

6.2 Performance Comparison

The zero-shot hallucination detection results with four popular LLMs are illustrated in Tab. 3. With different LLMs, similar comparison conclusions

[2] https://huggingface.co/meta-llama/Llama-2-7b
[3] https://huggingface.co/meta-llama/Llama-2-13b
[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[5] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

Approaches	Llama2-7b			Llama2-13b		
	Books	Movies	GCI	Books	Movies	GCI
SBERT	0.459	0.519	0.957	0.573	0.539	0.960
SelfCheckGPT	0.783	0.811	0.790	0.751	0.794	0.885
INSIDE	0.776	0.832	0.837	0.771	0.811	0.913
InterrogateLLM	0.819	0.891	0.961	0.804	0.842	0.966
AGSER	**0.859**	**0.935**	**0.974**	**0.810**	**0.884**	**0.988**
Approaches	Llama3-8b			Qwen2.5-14b		
	Books	Movies	GCI	Books	Movies	GCI
SBERT	0.763	0.639	0.969	0.573	0.626	0.505
SelfCheckGPT	0.825	0.802	0.721	0.711	0.763	0.607
INSIDE	0.846	0.791	0.766	0.703	0.751	0.667
InterrogateLLM	0.881	0.839	**0.990**	0.758	0.798	0.735
AGSER	**0.895**	**0.852**	0.986	**0.776**	**0.860**	**0.808**

Table 3: Performance comparison on zero-shot hallucination detection in LLMs.

can be observed. Not surprisingly, SBERT performs poorly, for it has no special design for measuring hallucinations in LLMs. Detecting hallucinations in output space and embedding space respectively, SelfCheckGPT and INSIDE have similar detection results. With detection AUC about 80%, they show their effectiveness in hallucination detection. Meanwhile, via asking reverse questions, InterrogateLLM improves the detection results by large margins. It allows the LLMs to rethink the generated answers from a new perspective, rather than only conducting multiple response resampling. Moreover, obviously, compared to the above state-of-the-art approaches, our proposed AGSER approach achieves the best hallucination detection results. With Llama2-7b, AGSER improves SelfCheckGPT, INSIDE and InterrogateLLM by 16.1%, 13.2% and 3.6% in average, respectively. With Llama2-13b, AGSER improves SelfCheckGPT, INSIDE and InterrogateLLM by 10.4%, 7.5% and 2.8% in average, respectively. With Llama3-8b, AGSER improves SelfCheckGPT, INSIDE and InterrogateLLM by 16.4%, 13.7% and 0.9% in average, respectively. With Qwen2.5-14b, AGSER improves SelfCheckGPT, INSIDE and InterrogateLLM by 17.4%, 15.2% and 6.7% in average, respectively. AGSER can significantly improve the detection performance with different LLMs across different datasets. The only exception is evaluating with Llama3-8b on the GCI dataset, in which the detection AUC is nearly 1.0. These observations strongly demonstrate the superiority of using attention values to guide LLMs to conduct self-reflection for detecting hallucinations.

6.3 Ablation Study

To investigate the effects of components and options in our proposed AGSER approach, we perform extensive ablation studies, and report the corresponding results.

Firstly, we investigate the effects of attentive and non-attentive queries, respectively. Hallucination detection results of AGSER with only attentive queries or non-attentive queries are shown and compared to the results of AGSER in Tab. 4. Obviously, attentive query plays the major role in the effectiveness of AGSER. And AGSER with only non-attentive queries achieves hallucination detection AUC of 0.575 in average, which indicates non-attentive queries are also necessary for hallucination detection. Specifically, without consideration of attentive queries, the detection AUC of AGSER decreases by 38.6%, 33.3%, 40.7% and 26.6% in average with the four LLMs respectively. Meanwhile, without consideration of non-attentive queries, the detection AUC of AGSER decreases by 0.9%, 0.4%, 0.6% and 1.4% in average with the four LLMs respectively. The above observations are reasonable, because only in a small portion of hallucination samples, the answers of non-attentive queries shall stay unchanged. It is not an extremely strong indicator, but still a necessary one for reflecting the reasoning and answer generating process in LLMs. In a word, both attentive and non-attentive queries are necessary and effective for detecting hallucinations in LLMs.

Approaches	Llama2-7b			Llama2-13b		
	Books	Movies	GCI	Books	Movies	GCI
AGSER	**0.859**	**0.935**	**0.974**	0.810	**0.884**	**0.988**
AGSER w/ attentive queries	0.848	0.926	0.970	**0.814**	0.875	0.984
AGSER w/ non-attentive queries	0.572	0.581	0.545	0.508	0.649	0.631
Approaches	Llama3-8b			Qwen2.5-14b		
	Books	Movies	GCI	Books	Movies	GCI
AGSER	**0.895**	**0.852**	**0.986**	**0.776**	**0.860**	**0.808**
AGSER w/ attentive queries	0.887	0.846	0.984	0.765	0.846	0.800
AGSER w/ non-attentive queries	0.553	0.556	0.511	0.581	0.625	0.589

Table 4: Ablation study results regarding using only attentive or non-attentive queries for hallucination detection.

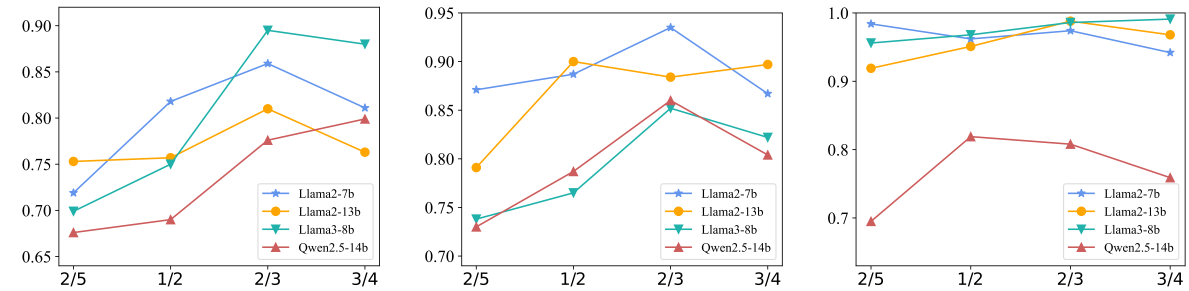
Approaches	Llama2-7b			Llama2-13b		
	Books	Movies	GCI	Books	Movies	GCI
AGSER w/ s_i^{first}	0.746	0.909	0.883	0.686	0.878	0.831
AGSER w/ s_i^{mid}	0.771	0.884	**0.974**	0.771	**0.889**	0.954
AGSER w/ s_i^{last}	0.792	0.849	0.962	0.741	0.815	0.973
AGSER w/ s_i^{max}	0.801	0.932	0.923	0.717	0.855	0.903
AGSER w/ s_i^{mean}	**0.859**	**0.935**	**0.974**	**0.810**	0.884	**0.988**
Approaches	Llama3-8b			Qwen2.5-14b		
	Books	Movies	GCI	Books	Movies	GCI
AGSER w/ s_i^{first}	0.727	0.790	0.862	0.669	0.779	0.765
AGSER w/ s_i^{mid}	0.848	0.843	0.941	0.676	**0.882**	0.761
AGSER w/ s_i^{last}	0.709	0.847	0.837	0.699	0.843	0.793
AGSER w/ s_i^{max}	0.753	0.815	0.979	0.756	0.836	0.762
AGSER w/ s_i^{mean}	**0.895**	**0.852**	**0.986**	**0.776**	0.860	**0.808**

Table 5: Ablation study results regarding different token contribution scores.

Secondly, we investigate the effects of different token contribution scores. As introduced in Sec. 5, there are five different token contribution scores: s_i^{first}, s_i^{mid}, s_i^{last}, s_i^{max} and s_i^{mean}. Accordingly, we report the hallucination detection results of AGSER with s_i^{first}, s_i^{mid}, s_i^{last}, s_i^{max} and s_i^{mean} respectively in Tab. 5. AGSER with s_i^{first} achieves the lowest detection AUC of only 0.794 in average. Only considering the first layer attention contributions, we may lose some important states in the latter layers. Considering the attention contributions in the last layer, which integrate some useful states in the formal layers, AGSER with s_i^{last} achieves better detection AUC of 0.822 in average. Meanwhile, using the attention contributions in the middle layer, AGSER with s_i^{mid} further improves the hallucination detection AUC to 0.849 in average. Moreover, with max pooling and mean pooling, we can capture the overall characteristics of all layers in LLMs more comprehensively, and thus achieve satisfactory hallucination detection results. AGSER with s_i^{max} and s_i^{mean} achieves detection AUC of 0.836 and 0.886 in average, respectively. Using the maximum values of all layers is obviously worse, indicating that max pooling may neglect some important information across different layers in LLMs. Meanwhile, using the mean values of all layers is clearly better, and s_i^{mean} is the best token contribution score according to our experimental results.

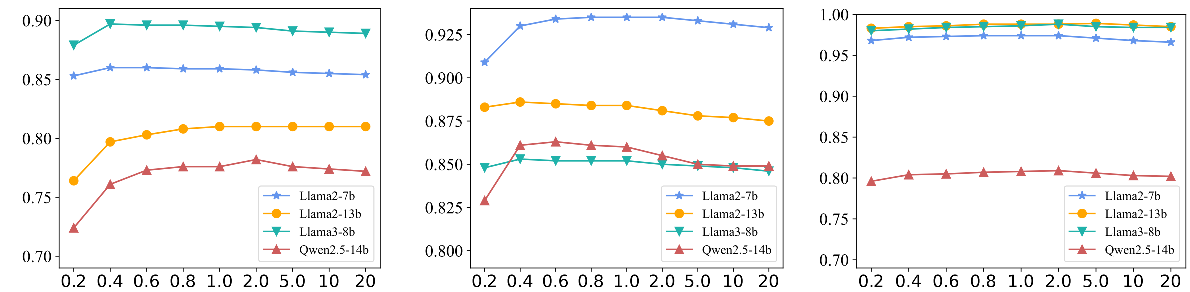
6.4 Hyper-parameter Study

To investigate the impact of hyper-parameters in AGSER on the hallucination detection results, we conduct some hyper-parameter studies. Firstly, we show the detection AUC with varying k values in



(a) The Books dataset. (b) The Movies dataset. (c) The GCI dataset.

Figure 2: Hallucination detection results evaluated by AUC with varying k values.



(a) The Books dataset. (b) The Movies dataset. (c) The GCI dataset.

Figure 3: Hallucination detection results evaluated by AUC with varying λ values.

Fig. 2. The hyper-parameter k controls the percentage of tokens selected for the attentive query. In general, with larger k values, which means retaining more sufficient major information in attentive queries, the results tend to be better. But when $k = 3/4$, in some cases, the detection results decrease slightly. Secondly, we show the detection AUC with varying λ values in Fig. 3. The hyper-parameter λ controls the balance between attentive and non-attentive consistency scores. In general, with different λ values, the results are relatively stable. Meanwhile, focusing too much on attentive or non-attentive consistency scores, AGSER will show some performance decline.

6.5 Discussions

According to the above observations, AGSER significantly outperforms state-of-the-art approaches on zero-shot hallucination detection in LLMs. In addition, AGSER requires a lower computational overhead of resampling. The compared methods, i.e., SelfCheckGPT, INSIDE and InterrogateLLM, perform 5 times of LLM running. In contrast, AGSER only requires 3 times of LLM running (feeding original, attentive and non-attentive queries into LLMs), and 2 times of token usage (attentive and non-attentive queries together have the same tokens as the original one). In a word, AGSER has great advantages in both effectiveness and efficiency. Furthermore, some running example results and bad cases of AGSER are presented in Apps. H and I respectively.

7 Conclusion

In summary, this work presents a systematic investigation of attention mechanisms in LLMs and proposes AGSER, a novel and computationally efficient approach for zero-shot hallucination detection. Through extensive experiments on three distinct factual knowledge recall tasks with four widely-used LLMs, AGSER demonstrates superior performance compared to existing hallucination detection methods. Our findings make several key contributions to the field: (1) we provide new insights into how attention patterns correlate with hallucination behaviors in LLMs; (2) we establish AGSER as a robust and resource-efficient framework for hallucination detection. We believe that this work represents a significant step toward more reliable and trustworthy large language models.

Limitations

While AGSER demonstrates promising results, we acknowledge several limitations of our approach.

First, the method's reliance on attention allocation patterns during inference restricts its applicability to open-source LLMs, making it challenging to detect hallucinations in closed-source models accessed through APIs.

Furthermore, while AGSER achieves a remarkable 50% or greater reduction in computational overhead compared to existing self-consistency methods, representing a significant breakthrough in efficiency, our approach still requires three inference passes with two token sets. The remaining computational requirements may still present challenges in specific scenarios, such as real-time applications or resource-constrained environments.

Ethical Considerations

While our work aims to detect hallucinations, it is crucial to note that LLMs may still produce unreliable, biased, or factually incorrect information. Therefore, we emphasize that the outputs from our experimental results should be interpreted primarily as indicators of hallucination detection effectiveness rather than as reliable sources of factual information.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Conference on Empirical Methods in Natural Language Processing*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. In *Conference on Empirical Methods in Natural Language Processing*.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, et al. 2025. Integrative decoding: Improve factuality via implicit self-consistency. In *International Conference on Learning Representations*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024a. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Conference on Empirical Methods in Natural Language Processing*.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. 2025. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling. In *AAAI Conference on Artificial Intelligence*.

Jinwen He, Yujia Gong, Zijin Lin, Yue Zhao, Kai Chen, et al. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of ACL*.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*.

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

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. 2024. Enhancing language model factuality via activation-based confidence calibration and guided decoding. *arXiv preprint arXiv:2406.13230*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Team Qwen. 2024. Qwen2.5: A party of foundation models.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. Interrogatellm: Zero-resource hallucination detection in llm-generated answers. In *Annual Meeting of the Association for Computational Linguistics*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023b. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of EMNLP*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Annual Meeting of the Association for Computational Linguistics*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Details of Datasets

We show the statistics of the Books, Movies and GCI datasets respectively in Tab. 6. In this work, as we aim to investigate the problem of zero-shot hallucination detection in LLMs, we use all the

	Books	Movies	GCI
Number of Samples	3000	3000	181

Table 6: Statistics of the datasets.

samples in the datasets for testing, and there are no training samples.

B More Pilot Study Results

Following the analysis in Sec. 4, in this section, we present more pilot study results. We provide more results with Llama2-7b, Llama2-13b, Llama3-8b and Qwen2.5-14b on the Books, Movies and GCI datasets. The corresponding results are shown in Tabs. 7-28. We can draw the same conclusion as in Sec. 4, i.e., smaller attentive consistency scores and larger non-attentive consistency scores indicate greater probabilities of hallucinations in LLMs.

C More Baseline Introduction

The compared zero-shot hallucination detection approaches are introduced as follows:

- **SBERT**: Following (Yehuda et al., 2024), we employ a pre-trained Sentence BERT model (Reimers and Gurevych, 2019) as a baseline, which embeds both query and answer into vectors. Then, we calculate the cosine similarity between them as the hallucination prediction.
- **SelfCheckGPT** (Manakul et al., 2023): A detection approach that generates multiple responses and verifies whether they support the original answer.
- **INSIDE** (Chen et al., 2024): An approach that calculates eigenvalues of multiple answers in the sentence embedding space as the hallucination prediction estimator.
- **InterrogateLLM** (Yehuda et al., 2024): A state-of-the-art approach that detects hallucinations via feeding the reverse question into LLMs and verifies whether the original query could be generated.

D More Detailed Settings

The LLMs used in our experiments are introduced as follows:

- **Llama 2-7B** is a variant of the Llama 2 family, and released in July 2023. It features 7 billion parameters, and is designed to perform a variety of natural language processing tasks.
- **Llama 2-13B** is also a variant of the Llama 2 family, and released in July 2023. It features 13 billion parameters.
- **Llama 3-8B** is a LLM from the Llama 3 series. It features 8 billion parameters, and is released in April 2024. It is one of the most advanced open-source LLMs.
- **Qwen 2.5-14B** is a LLM from the Qwen series. Released in September 2024, this model features 14 billion parameters. It is also one of the most advanced open-source LLMs, and shows great Chinese ability.

Moreover, all experiments are conducted on NVIDIA A100 GPUs with 80GB of memory. We utilize a fixed random seed of 42, and the experimental results are reported within a single run. Meanwhile, in our experiments, we employ the following versions of the libraries and models: SpaCy version 2.3.9, transformers version 4.30.2, and rouge version 1.0.1.

E Licensing

The Books, Movies and GCI datasets are released for academic usage. These datasets are designed for hallucination detection. Thus, our use of these datasets is consistent with their intended use.

Moreover, Llama 2-7B and Llama 2-13B are released under the Meta Llama 2 Community License Agreement. Llama 3-8B is released under the Meta Llama 3 Community License Agreement. And Qwen 2.5-14B is released under the Apache-2.0 License. They are all open for academic usage.

F Prompts

In this section, we detail the prompts for generating answers in LLMs. The prompt template is shown in Fig. 4. And example prompts in the Books, Movies and GCI datasets are illustrated in Figs. 5-7 respectively.

G More Ablation Study Results

In addition to the token contribution scores discussed in Sec. 5, we investigate more layers in LLMs for token contribution calculation. Results with different LLMs are shown in Tabs. 29-32. We can see that, AGSER w/ s_i^{mean} can achieve the best overall performances. And using values in some specific layers for calculating the token contribution scores can result in relatively high detection results in minor cases.

H Example Results

In this section, we present some running example results of AGSER in Tabs. 33-40. We can observe that, for non-hallucination samples, compared to the original answers, the answers of the attentive queries stay consistent, and those of the non-attentive queries otherwise. And for hallucination samples, the answers of the attentive queries mostly change, and those of the non-attentive queries may remain unchanged. These observations enable our proposed AGSER approach to accurately detect hallucinations in LLMs.

I Bad Cases

To investigate the shortage of AGSER and potential improvement, we demonstrate some bad cases of AGSER:

- For the query “Who is the author of the book Nights in Rodanthe, what year was it published?”, the LLM correctly responded with “Nicholas Sparks, in 2002.” However, the attentive query was incorrectly segmented as “Riding the Bus with My Sister: A True Life Journey, what year?”, omitting the request for the author’s name. Consequently, the LLM only answered “In 2002,” resulting in a final attentive consistency score of just 0.4 for this non-hallucination sample.

- Regarding the question “Who is the author of the book Who Moved My Cheese?, what year was it published?”, the LLM erroneously answered “Spencer Johnson, in 1996” (the correct publication year being 1998). When the same question was posed as an attentive query, the response remained “Spencer Johnson, in 1996,” leading to an attentive consistency score of 0.99 for this hallucination sample. This indicates that the LLM maintains incorrect memories about less commonly referenced information (such as book publication years).

- For the query “What actors played in the 1944 movie House of Frankenstein?”, the LLM initially provided the correct answer: “The main cast included Boris Karloff, J. Carrol Naish and Lon Chaney Jr.” However, the attentive query was mistakenly segmented as “What actors played in the 1944 movie?”, omitting the movie title. This led the LLM to incorrectly respond with “Peter Lorre,” an actor active in the 1940s, resulting in an attentive consistency score of only 0.24 for this non-hallucination sample.

Based on these bad cases, we can conclude that AGSER’s erroneous judgments primarily stem from either incorrect segmentation of attentive queries (leading to omission of key information) or the LLM’s inherent memory inaccuracies (especially for less commonly referenced information). These observations will help us further optimize our detection methods and develop more robust query segmentation strategies in future work.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.092	0.130	0.212	0.566
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.610	0.210	0.102	0.078

Table 7: Distribution of attentive consistency scores r^{att} with Llama2-13b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.989	0.011	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.789	0.186	0.022	0.003

Table 8: Distribution of non-attentive consistency scores r^{non_att} with Llama2-13b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.0	0.0	0.432	0.568
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.822	0.108	0.007	0.063

Table 9: Distribution of attentive consistency scores r^{att} with Llama3-8b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.986	0.012	0.001	0.001

Table 10: Distribution of non-attentive consistency scores r^{non_att} with Llama3-8b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.127	0.181	0.262	0.430
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.722	0.114	0.053	0.111

Table 11: Distribution of attentive consistency scores r^{att} with Qwen2.5-14b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.987	0.013	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.907	0.070	0.015	0.008

Table 12: Distribution of non-attentive consistency scores r^{non_att} with Qwen2.5-14b on the Books dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.051	0.165	0.189	0.595
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.456	0.430	0.103	0.011

Table 13: Distribution of attentive consistency scores r^{att} with Llama2-7b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.975	0.023	0.001	0.001

Table 14: Distribution of non-attentive consistency scores r^{non_att} with Llama2-7b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.026	0.117	0.320	0.537
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.330	0.434	0.219	0.017

Table 15: Distribution of attentive consistency scores r^{att} with Llama2-13b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.864	0.128	0.007	0.001

Table 16: Distribution of non-attentive consistency scores r^{non_att} with Llama2-13b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.064	0.165	0.222	0.549
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.442	0.357	0.192	0.009

Table 17: Distribution of attentive consistency scores r^{att} with Llama3-8b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.994	0.004	0.001	0.001

Table 18: Distribution of non-attentive consistency scores r^{non_att} with Llama3-8b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.121	0.152	0.303	0.424
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.670	0.294	0.032	0.004

Table 19: Distribution of attentive consistency scores r^{att} with Qwen2.5-14b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.917	0.079	0.003	0.001

Table 20: Distribution of non-attentive consistency scores r^{non_att} with Qwen2.5-14b on the Movies dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.0	0.0	0.013	0.987
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.962	0.038	0.0	0.0

Table 21: Distribution of attentive consistency scores r^{att} with Llama2-7b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.990	0.010	0.00	0.0

Table 22: Distribution of non-attentive consistency scores r^{non_att} with Llama2-7b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.0	0.0	0.080	0.920
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.993	0.007	0.0	0.0

Table 23: Distribution of attentive consistency scores r^{att} with Llama2-13b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.840	0.120	0.020	0.020

Table 24: Distribution of non-attentive consistency scores r^{non_att} with Llama2-13b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.0	0.0	0.025	0.975
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.986	0.014	0.0	0.0

Table 25: Distribution of attentive consistency scores r^{att} with Llama3-8b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
1.0	0.0	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.936	0.043	0.021	0.0

Table 26: Distribution of non-attentive consistency scores r^{non_att} with Llama3-8b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.011	0.011	0.024	0.954
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.818	0.152	0.030	0.0

Table 27: Distribution of attentive consistency scores r^{att} with Qwen2.5-14b on the GCI dataset.

Non-hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.994	0.006	0.0	0.0
Hallucination Samples			
[0.0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1.0]
0.894	0.061	0.030	0.015

Table 28: Distribution of non-attentive consistency scores r^{non_att} with Qwen2.5-14b on the GCI dataset.

Prompts

You are a helpful intelligent chatbot to answer questions.
Follow the format below, and please only predict the answer that corresponds to the last question.
Question: {question}
Answer:

Figure 4: Prompts to answer the questions.

Prompts

You are a helpful intelligent chatbot to answer questions.
Follow the format below, and please only predict the answer that corresponds to the last question.
Question: Who is the author of the book Classical Mythology, what year was it published?
Answer:

Figure 5: Example prompts in the Books dataset.

Layer	Books	Movies	GCI
8	0.789	0.888	0.969
24	0.801	0.877	0.962

Table 29: More ablation study results with Llama2-7b.

Prompts

You are a helpful intelligent chatbot to answer questions.
Follow the format below, and please only predict the answer that corresponds to the last question.
Question: What actors played in the 1995 movie Jumanji?
Answer:

Figure 6: Example prompts in the Movies dataset.

Prompts

You are a helpful intelligent chatbot to answer questions.
Follow the format below, and please only predict the answer that corresponds to the last question.
Question: What is the capital of France?
Answer:

Figure 7: Example prompts in the GCI dataset.

Layer	Books	Movies	GCI
10	0.784	0.868	0.961
30	0.771	0.836	0.959

Table 30: More ablation study results with Llama2-13b.

Layer	Books	Movies	GCI
8	0.803	0.842	0.986
24	0.744	0.857	0.996

Table 31: More ablation study results with Llama3-8b.

Layer	Books	Movies	GCI
8	0.719	0.827	0.728
16	0.757	0.790	0.768
32	0.729	0.784	0.826
40	0.692	0.787	0.695

Table 32: More ablation study results with Qwen2.5-14b.

Query	Who is the author of the book Dreamcatcher, what year was it published?
Correct Answer	Stephen King, in 2001.
Original Answer	Stephen King, in 2001.
Attentive Query	author book Dreamcatcher, year it published?
Answer	Stephen King, in 2001.
Non-attentive Query	Who is the of the what was
Answer	Carlo D'Este.

Table 33: AGSER's running example result 1.

Query	Who is the author of the book Summer Sisters, what year was it published?
Correct Answer	Judy Blume, in 1998.
Original Answer	Judy Blume, in 1998.
Attentive Query	book Summer Sisters, what it published?
Answer	Judy Blume, in 1998.
Non-attentive Query	Who is the author of the year was
Answer	Mark P. O. Morford.

Table 34: AGSER's running example result 2.

Query	Who is the author of the book Coraline, what year was it published?
Correct Answer	Neil Gaiman, in 2002.
Original Answer	Neil Gaiman, in 2002.
Attentive Query	author book Coraline, year it published?
Answer	Neil Gaiman, in 2002.
Non-attentive Query	Who is the of the what was
Answer	Carlo D'Este.

Table 35: AGSER's running example result 3.

Query	Who is the author of the book At Home in Mitford, what year was it published?
Correct Answer	Jan Karon, in 1996.
Original Answer	Jan Karon, in 1996.
Attentive Query	the author of the At Home in Mitford, was it published?
Answer	Jan Karon, in 1996.
Non-attentive Query	Who is book The what year
Answer	The author of The Nightingale is Kristin Hannah, and it was published in 2015.

Table 36: AGSER's running example result 4.

Query	Who is the author of the book Final Stand, what year was it published?
Correct Answer	Helen Myers, in 2002.
Original Answer	Mark P. O. Morford.
Attentive Query	author of book Final Stand, what year it published?
Answer	Michael Stephenson, in 2007.
Non-attentive Query	Who is the the was
Answer	Mark P. O. Morford.

Table 37: AGSER's running example result 5.

Query	Who is the author of the book Secrets of St. John's Wort: A Lynn Sonberg Book, what year was it published?
Correct Answer	Larry Katzenstein, in 1998.
Original Answer	Lynn Sonberg, in 2003.
Attentive Query	. John's Wort: A Lynn Sonberg Book,?
Answer	2001.
Non-attentive Query	Who is the author of the book Secrets of St what year was it published
Answer	Mary's Hospital, in 2003.

Table 38: AGSER's running example result 6.

Query	Who is the author of the book My Cat Spit McGee, what year was it published?
Correct Answer	Willie Morris, in 1999.
Original Answer	Mark P. O. Morford, in 2002.
Attentive Query	author book My Cat Spit McGee, published?
Answer	Iain Levison, in 2004.
Non-attentive Query	Who is the of the what year was it
Answer	Mark P. O. Morford, in 2002.

Table 39: AGSER's running example result 7.

Query	Who is the author of the book Secrets of St. John's Wort: A Lynn Sonberg Book, what year was it published?
Correct Answer	Marshall Kirk, in 1989.
Original Answer	1990
Attentive Query	book After Ball: Americaquerays in '90s, what year it published?
Answer	1999
Non-attentive Query	Who is the author of the the How Will Con Its Fear and Hatred of G the was
Answer	Thomas Pynchon, in 1990.

Table 40: AGSER's running example result 8.