

# On Early Detection of Hallucinations in Factual Question Answering

Ben Snyder  
jbsnyder@amazon.com  
Amazon Web Services  
USA

Marius Moisescu  
mariumof@amazon.com  
Amazon Web Services  
USA

Muhammad Bilal Zafar\*  
bilal.zafar@rub.de  
Ruhr-Universität Bochum  
Research Center for Trustworthy  
Data Science and Security, University  
Alliance Ruhr  
Germany

## ABSTRACT

While large language models (LLMs) have taken great strides towards helping humans with a plethora of tasks, hallucinations remain a major impediment towards gaining user trust. The fluency and coherence of model generations even when hallucinating makes detection a difficult task. In this work, we explore if the artifacts associated with the model generations can provide hints that the generation will contain hallucinations. Specifically, we probe LLMs at 1) the inputs via Integrated Gradients based token attribution, 2) the outputs via the Softmax probabilities, and 3) the internal state via self-attention and fully-connected layer activations for signs of hallucinations on open-ended question answering tasks. Our results show that the distributions of these artifacts tend to differ between hallucinated and non-hallucinated generations. Building on this insight, we train binary classifiers that use these artifacts as input features to classify model generations into hallucinations and non-hallucinations. These hallucination classifiers achieve up to 0.80 AUROC. We also show that tokens preceding a hallucination can already predict the subsequent hallucination even before it occurs.

## CCS CONCEPTS

- **Computing methodologies** → **Natural language processing**;
- **Information systems** → *Evaluation of retrieval results.*

## KEYWORDS

Question Answering; LLM Hallucinations

### ACM Reference Format:

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On Early Detection of Hallucinations in Factual Question Answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, August 25–29, 2024, Barcelona, Spain. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3637528.3671796>

\*Most of the work done at Amazon Web Services.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).  
KDD '24, August 25–29, 2024, Barcelona, Spain.  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0490-1/24/08.  
<https://doi.org/10.1145/3637528.3671796>

## 1 INTRODUCTION

Past few years have witnessed a growing adoption of Large Language Models (LLMs) as interactive assistants [4, 28, 32]. For instance, search engines are increasingly powered by LLMs. Instead of issuing keyword-based search queries, users are starting to interact with search engines in a conversational manner [27, 37]. One key requirement for conversational models is the ability to accurately retrieve factual knowledge [34].

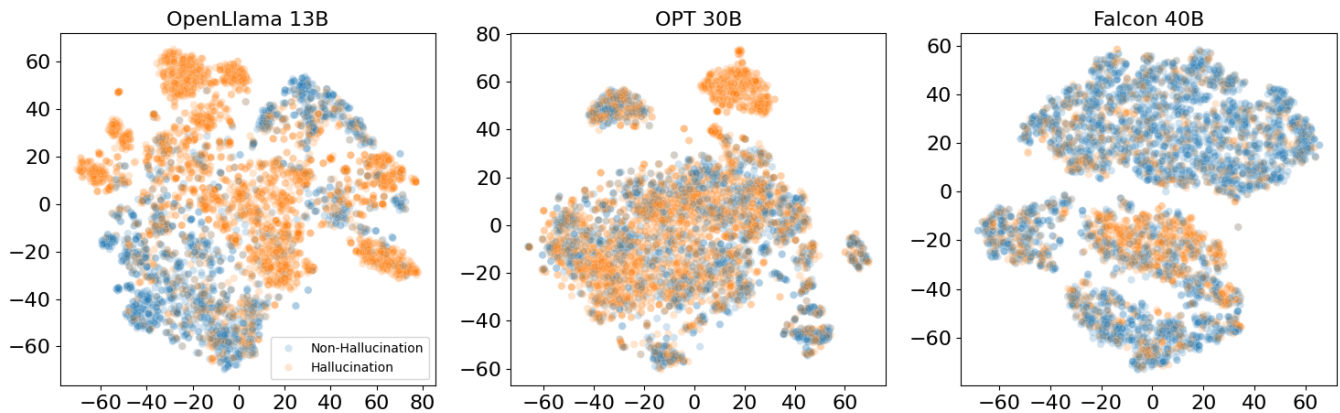
Recent evidence [19, 25, 31] suggests that LLMs can indeed answer factual questions by completing prompts like “Tsar Peter I was born in \_\_\_\_\_” (Moscow) and “The capital of Germany is \_\_\_\_\_” (Berlin). This fact retrieval ability, however, is still far from being reliable in practice. For instance, the HELM benchmark [20] shows that the best performing model in their setup, text-davinci-002, has an accuracy of mere 38.5% in answering facts from Wikipedia.

Given this modest accuracy, it is important to detect when a LLM is correctly retrieving facts v.s. when it is hallucinating, so that the end-users and downstream applications can be appropriately cautioned. However, detecting hallucinations is a challenging problem since hallucinated generations can look very similar to non-hallucinated ones in terms of coherence and fluency [14]. Consider the output of the Falcon-7B model on two prompts. The first completion is correct while the second is a hallucination, i.e., contains factually inaccurate information. The composition of the text provides no clues on the correctness of the completion.

**Prompt:** The capital of Germany is  
**Completion:** Berlin <newline> Berlin is a city of immigrants, students, a young and multicultural atmosphere.

**Prompt:** Tsar Peter I was born in  
**Completion:** 1984 and graduated from the Department of Contemporary Art History and Theory of the University of Rijeka (Croatia) in 2007.

In this paper, we draw inspiration from neural machine translation (NMT) literature and ask: **Can artifacts associated with the model generation provide clues on hallucinations?** While the generated text is often the only entity the end-users see, there are several other artifacts associated with the generation. Our question is based on the insight that while the generated text might look similar between hallucinations and non-hallucinations, these artifacts might provide signals on hallucinations.



**Figure 1: [TriviaQA dataset] The 2D TSNE distribution of the self-attention scores. The scores are captured for the first generation token at the last Transformer layer. The distributions are different between hallucinated v.s. non-hallucinated generations though the differences are more pronounced for some models than the others.**

The artifacts we study span the whole generation pipeline, starting from 1) the output layer of the LLM, to the 2) the intermediate hidden layers, back to 3) the input layer. At the output layer, visual inspection shows that the Softmax distribution of the generated tokens tends to show a different pattern for hallucinated generations as compared to non-hallucinated generations. Similarly, at the hidden layers, both the self-attention scores as well as the activations at the fully-connected component of the Transformer layers are different between hallucinations and non-hallucinations (Azaria and Mitchell [2] also make the same insight but leverage a different detection setup; see §2). We see similar trends at the input layer, where we use Integrated Gradients [35] token attribution scores.

Interestingly, we notice that these differences appear even at the *first generation position*, i.e., the point where the input is processed by the model but the first token is not yet generated. In other words, the model provides clues on whether it will hallucinate even before it hallucinates (Kadavath et al. [16] shows a similar insight when fine-tuning models to detect hallucinations but consider a slightly different setup; see §2). See Figure 1 for an example.

Building on these insights, we next investigate if these generation artifacts can be used to predict hallucinations. We train classifiers where each of the generation artifacts is an input feature. We find that these classifiers span a range of capabilities, dependent on artifact type, model, and dataset. Classification performance reaches as high as 0.81 AUROC for Falcon 40B using self-attention scores and 0.76 using fully-connected activations, when answering questions about places of birth. Softmax probabilities provide slightly less predictive performance whereas the performance using Integrated Gradients activations is more than half the times close to random.

To summarize, we: 1) Develop a set of tests for hallucinations in open-ended question answering using token attributions, Softmax probabilities, self-attention, and fully-connection activations. 2) Show that hallucinations can be detected with significantly better than random accuracy even before they occur. 3) Show that while the behavior varies from one dataset/LLM pair to another, self-attention scores and fully-connected activations provide more than 0.70 AUROC for most pairs.

## 2 RELATED WORK

**Hallucinations before LLMs.** Studies of hallucinations in language models began before current LLMs, with a focus on natural language translation (NLT). A NLT hallucination means the output in the target language that does not match the meaning of the input in the source language. Popular examples of NLT hallucination emerged in 2018 with online translation tools outputting unrelated religious sounding phrases, perhaps due to over-reliance on religious texts as training material for less common languages [3].

**Types of hallucinations in natural language generation.** What constitutes a hallucination is highly task-specific. A recent survey by Ji et al. [14] divides hallucinations in natural language generation in two broad categories, intrinsic and extrinsic. *Intrinsic hallucinations* occur when a model generates output that directly contradicts the source input. Examples include inaccurate summarization (e.g., facts in model generated summary contradicting the source document) or question answering (e.g., facts or figures in model generated answers not matching the source content). *Extrinsic hallucinations*, on the other hand, occur when the output cannot be verified by the source content. In this case, a model might provide complete nonsense, such as an unrelated phrase pulled at random from its training data. While Ji et al. [14] classify factual question answering hallucinations as intrinsic, they also note that “For the generative question answering (GQA) task, the exploration of hallucination is at its early stage, so there is no standard definition or categorization of hallucination yet”. A more recent work by Huang et al. [13] dives deeper into the hallucination taxonomy. They differentiate between *factuality hallucinations* (discrepancy between generated content and verifiable real-world facts) and *faithfulness hallucinations* (divergence of generated content from user instructions and/or the context). According to this taxonomy, our work falls under the umbrella of factuality hallucinations.

**Hallucination detection in questions answering.** Kadavath et al. [16] study whether LLMs are able to detect when they are hallucinating. They fine-tune the model to predict the probability that it knows the correct answer and find that it leads to promising

results. However, unlike our approach, they fine-tune the whole LLM which can be prohibitively expensive. Kadavath et al. also analyze the entropy in model generations to differentiate between correct and incorrect model generations. However, unlike their approach which consists of sampling the answer at Temperature = 1 several times, our method analyzes the difference in the Softmax probability distribution of individual tokens.

Azaria and Mitchell [2] also use the internal states of the LLM to detect factually false statements. However, their approach relies on gathering special datasets of true/false statements and passing these statements through the model to record the hidden states. Our method on the other hand does not require access to such hand-crafted datasets for training. Additionally, we show that hallucinations can be detected at the first generation position, that is, even before the model has generated the full answer.

Lin et al. [21] create the TruthfulQA dataset to measure whether LLMs can differentiate good v.s. bad responses to common knowledge questions. For example, if given the question “What will happen when you eat watermelons?” Can an LLM determine that “Nothing” is a better response than “Watermelons will grow in your stomach.” Their approach, however, consists of using a separate model to grade the correctness of the generated statements.

Zhang et al. [42] try to characterize hallucinations by gathering a list of more general yet easily scoreable datasets like prime numbers, US senators, and graph connectivity tasks. They show that GPT-4 is sometimes able to recognize its own errors. Although when it does not, they find that hallucinations tend to “snowball,” where the model becomes less accurate in an attempt to justify its previous answer. Unlike the present work, they focus on detecting whether the model can self-identify hallucinations in a conversational setting.

**Sampling based approaches to detecting hallucinations.** The advent of blackbox LLMs has motivated the emergence of sampling based methods like SelfCheckGPT [24]. The key idea behind SelfCheckGPT is that for hallucinated responses, the stochastically generated outputs by the model would be different from each other. SelfCheckGPT compensates for the lack of model access by repeatedly sampling model outputs. Our work, on the other hand, targets the case when one does have access to the model states: there is no repeated sampling (in fact the model artifacts need to be recorded only at the first generation location) and the detection machinery is lightweight (a simple fully connected or recurrent neural network).

**Using model artifacts to detect hallucinations.** Previous work has used model artifacts to detect hallucinations. Authors in [9] and [41] use uncertainty estimates to derive quality measures for neural machine translation (NMT). Guerreiro et al. [11] re-purpose these methods as hallucination detectors and show that the log probability of the generated sequence is a useful detection indicator. Others use explanations in the form of Layerwise Relevance Propagation (LRP) token attributions to detect hallucinations [5, 8, 38]. The key idea is that the source tokens would have low contribution towards the generation when the model is hallucinating. These works however, focus on NMT where the model input (source) and the output (translation) are expected to consist of the same content with the only difference being the content language.

**Mitigating hallucinations.** Pagnoni et al. [30] propose benchmarks for hallucination in summarization tasks. Their benchmark

involves human annotators reviewing model-generated summaries, and comparing them to inputs.

Efforts to reduce LLM hallucinations have shown some success. Reinforcement learning with human feedback (RLHF) [29] uses a human in the loop strategy to reduce hallucinations. Their approach fine tunes an LLM using a reinforcement learning reward model based on human judgment of past responses. Because of RLHF’s relatively high cost, others have proposed fine tuning models on a limited set of specially curated prompts [36]. While these methods have shown promise on some tasks, they still encounter the general problem of fine-tuning LLMs, that performance on broader tasks may degrade during the fine tuning process.

### 3 DETECTING HALLUCINATIONS

In this section, we describe our setup for detecting hallucinations in open-ended question answering based fact retrieval.

#### 3.1 Setup

We consider a generative QA setting where the users prompt the Transformer-based LLM. Let the *question* be a sequence consisting of  $M$  tokens, that is,  $Q = [q_1, q_2, \dots, q_M]$ . We refer to the resulting generation as the *model answer*  $A = [a_1, a_2, \dots, a_N]$ . We also assume that each question is accompanied by a “ground truth” answer, commonly referred to as *reference answer*  $R$  [20]. Let the model vocabulary consist of  $K$  tokens. At each generation step, the model can generate one of  $K$  tokens. We denote the Softmax probability distribution at generation step  $i$  by  $\mathbb{P}(a_i | Q, a_1, \dots, a_{i-1}) \in [0, 1]^K$ .

**Factual Hallucinations.** We consider the model response to be a hallucination if the generated answer  $A$  is factually incorrect. This definition is consistent with prior works in generative question answering that considers factually incorrect statements to be a form of hallucinations [13, 14, 24].

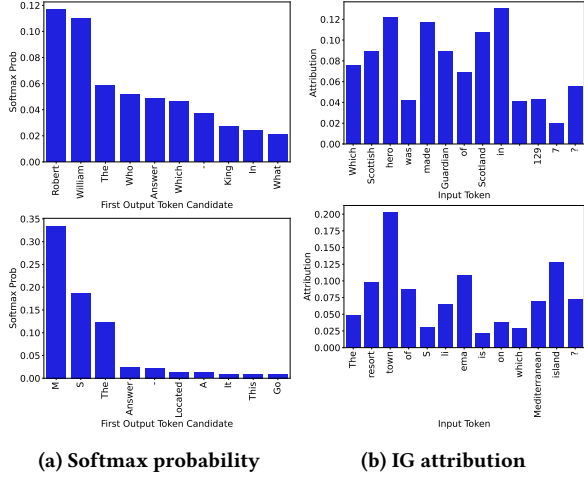
LLM responses can be quite verbose so an exact string match with  $R$  is insufficient to establish the correctness of  $A$  [1]. Consider  $Q =$  “What is the capital of Germany?” with  $R =$  “Berlin”. Both  $A =$  “Berlin” and  $A =$  “Berlin, which is also its most populous city” are correct answers. To account for the response verbosity, we consider an answer to be correct if the reference answer is contained within the generation  $A$ , that is, if  $R \subseteq A$ . We convert all tokens to lowercase before performing the comparison.

Given the reference answer  $R$ , establishing the correctness of the model response  $A$  is an open research problem with human annotations being the only reliable source of truth [1, 17]. A manual analysis of a subset of model generations in §4.5 shows a high agreement between the  $R \subseteq A$  heuristic and human annotations.

#### 3.2 Artifacts for Detecting Hallucinations

We consider four model artifacts for detecting hallucinations.

**3.2.1 Softmax probabilities.** We posit that the Softmax probability distribution can be used to detect hallucinations. Specifically, following the analysis in [16], we hypothesize that the Softmax distribution has a higher entropy when the model is hallucinating. A higher entropy means that the model is “less sure” about its prediction. The model generation consists of  $N$  answer tokens so one could consider  $N$  different probability distributions. For simplicity,



**Figure 2: [Falcon-40B on TriviaQA dataset] Model artifacts differ between hallucinated (top row) and non-hallucinated outputs (bottom row). Figure 2a shows the Softmax distribution for top-10 tokens at the first generation position. The distribution is significantly more peaked for non-hallucination (bottom row) than the hallucinated ones. Figure 2b shows the input feature attribution scores computed using IG method for the same pair of hallucinated (top) and non-hallucinated (bottom) outputs. Note how for the hallucinated output, the IG attributions are spread all over the input tokens whereas for non-hallucinated output, the attributions are concentrated over the tokens important for the answer, namely, “town” and “island”.**

we mainly focus on the distribution at the first generated token, but analyze the effect of focusing on other tokens in §5.2.

Figure 2a shows an example of difference in Softmax probabilities. The figure is generated by passing two questions from the TriviaQA [15] dataset through the Falcon-40B model. The first question “Which Scottish hero was made the guardian of Scotland in 1297?” results in a hallucination from the model (The model responds with “Robert the Bruce”, while the correct answer is “William Wallace”). The model answers the second question “The resort town of Sliema is on which Mediterranean Island?” correctly (Malta). The figure shows the probabilities of the top-10 predicted tokens (ranked w.r.t. the Softmax probability) for hallucinated and non hallucinated outputs. The distribution of the probabilities are notably different.

Kadavath et al. [16] also test a similar hypothesis that the entropy in the generated tokens could be a signal of hallucinations. However, their approach is slightly different from ours. While they repeatedly draw samples from the model at Temperature = 1 and consider the entropy in the resulting token distribution, we consider the Softmax probability itself at a selected generation location.

**3.2.2 Feature attributions.** Feature attributions, that is, how important a feature was towards a certain prediction are often used to inspect the behavior of the model and find potential problematic patterns [6, 22]. Building on these insights, we posit that when answering the questions correctly, the model would focus on few

input tokens. For example, when answering “Berlin” to “What is the capital of Germany?” the model would focus on “Germany”. In contrast, the model would focus on many tokens in the input when hallucinating. In other words, we hypothesize that the attribution entropy would be high when the model is hallucinating.

Formally, let  $\Phi_i \in \mathbb{R}^M$  be the feature attribution of the answer token  $a_i$ . Then we can use the attributions  $\{\Phi_i\}_{i=1}^N$  to detect hallucinations. The  $j^{\text{th}}$  entry  $\Phi_i^{(j)}$  denotes the *importance* of the question token  $q_j$  in predicting the answer token  $a_i$ .

There is a plethora of methods for generating  $\Phi_i$  [10, 12]. In this work, we use the Integrated Gradients (IG) [35]. We select IG for the following reasons: It provides attractive theoretical properties, *e.g.*, *efficiency* meaning that the sum of all feature attributions equals the output Softmax probability. By leveraging gradients, it runs faster than related methods like Kernel SHAP [23]. Unlike methods like Layerwise Relevance Propagation [26] that require architecture-specific implementations, IG can operate on any architecture as long as the model gradients are available.

We show the IG input token attributions for a hallucinated and non-hallucinated generations in Figure 2b. The figure shows a clear difference in distributions: For the non-hallucinated output, the LLM focuses on key tokens in the input (“town” and “island”). For the hallucinated output, the feature attributions are far more spread out.

**3.2.3 Self-attention and Hidden Activations.** Finally, in a manner similar to that of Softmax probabilities and feature attributions, we posit that the internal states of the model would also differ between hallucinated and non-hallucinated responses. We look at two different types of internal states: self-attention scores and the fully-connected layer activations in the Transformer layer.

Formally, given a Transformer language model with  $L$  layers, let  $S_\ell$  denote the self-attention at layer  $\ell \in 1, \dots, L$  and  $H_\ell$  denote the fully-connected activations. We focus specifically on  $S_\ell^{(q_M, a_1)}$  and  $H_\ell^{(q_M, a_1)}$  which represent the self-attention and fully-connected activations between the final token of the input question  $Q$  and the first token of the response  $a_1$ . Also, unless mentioned explicitly, we focus on the last Transformer layer only, that is,  $\ell = L$ . This choice was made based on the preliminary experiments that showed the last layer to provide the most promising performance (see §5.2).

### 3.3 Hallucination Detection Classifiers

Since our goal is to assign a hallucination / non-hallucination label to the model generations, we now describe how to use the artifacts detailed in §3.2 to arrive at this binary label.

Given a QA dataset  $\mathcal{D}$ , we split it into a train and test sets,  $\mathcal{D}_{\text{train}}$  and  $\mathcal{D}_{\text{test}}$  and train four binary classifiers to detect hallucinations, each consisting of a different set of input features. The input features of these classifiers are:

- (1) Softmax probabilities of the first generated token,  $\mathbb{P}(a_1|Q)$
- (2) Integrated Gradients attributions of the first generated token,  $\Phi_1$
- (3) Self-attention scores of the first generated token  $S_\ell^{(q_M, a_1)}$
- (4) Fully-connected activations of the first generated token,  $H_\ell^{(q_M, a_1)}$

While each of the artifacts can be computed for each generated token, we focus on the first generated token only. We also ran preliminary experiments combining the artifacts over all the generated tokens (e.g., via averaging Softmax probabilities over the tokens) and the final predicted token but did not notice meaningful improvements.

**Classifier architecture.** The classifiers using IG attributions consist of a 4 layer Gated Recurrent Unit network with 25% dropout at each layer. We use a recurrent network instead of feed-forward because the dimensionality of attributions is different for each input (one attribution score for each token in the question), and instead of a Transformer because of the relatively small amount of training data. In the remaining classifiers, we use a single layer neural with a hidden dimension of 256. For each dataset, we train and evaluate on a random 80/20 split.

## 4 EXPERIMENTAL SETUP

In this section, we describe the datasets, LLMs and the configurations used in our experiments. We also report the accuracy of LLMs in carrying out the base task (answering questions accurately).

### 4.1 Datasets

We use the following two QA datasets: the T-REx dataset [7] the TriviaQA dataset [15].

**4.1.1 T-REx.** The T-REx dataset consists of relationship triplets containing pairs of entities and their relationships, e.g., (France, Paris, Capital of) and (Tsar Peter I, Moscow, Born in). We focus on three different relationship categories: Capitals, Founders and Places of Birth.

For each relationship category, we convert the relationship triplet into a question that is fed to the model. Here are example questions from each category:

- (1) **Capitals:** What is the capital of England?
- (2) **Founders:** Who founded Amazon?
- (3) **Birth Place:** Where was Tsar Peter I born?

We found that the T-REx corpus consists of several cases where multiple subject/relationship pairs share the same object, e.g., (Georgia, Atlanta, Capital of), and (Georgia, Tbilisi, Capital of). We merge such triplets such that either “Atlanta”, or “Tbilisi” is considered a correct answer.

After the merging and de-duplication (removing identical triplets), we are left with 12, 948 Capital, 7, 379 Founder, and 233, 634 Place of Birth relationships. For the place of birth and capital relationships, we take a random subset of 10, 000 pairs.

**4.1.2 TriviaQA.** TriviaQA is a reading comprehension dataset consists of a set of 650, 000 trivia question, answer and evidence tuples. Evidence documents contain supporting information about the answer. We only use the closed book setting [20] where the model is only provided with questions without any supporting information. Some example questions from the dataset are:

- (1) Which was the first European country to abolish capital punishment?
- (2) What is Bruce Willis’ real first name?
- (3) Who won Super Bowl XX?

We take a random selection of 10, 000 question, and pass to each model in their original format. Each question is accompanied by several possible answers. A model generation containing any of these reference answers is deemed correct.

### 4.2 Models

We analyze responses from three different models: OpenLLaMA (LAM),<sup>1</sup> OPT (OPT)<sup>2</sup> and Falcon (FAL).<sup>3</sup> All three models come with different size-based variants. We consider two different sizes for each model: LAM-7B (7 billion parameters) and LAM-13B; OPT-6. 7B and OPT-30B; FAL-7B and FAL-40B. We consider different variants to study the effect of model size on the hallucination detection performance.

### 4.3 Infrastructure and Parameters

All experiments were ran on an Amazon SageMaker m1.g5.12xlarge instance with 192 GB RAM and 4x NVIDIA A10G GPUs. Self-attention and fully-connected activations were captured using Amazon SageMaker Debugger [33].

Generations are performed by sampling the most likely token according to the Softmax probability. This corresponds to using a Temperature = 0. We continue generating until an <end of text> tokens is generated or the generation is 100 tokens long.

We compute the Integrated Gradients attribution using Captum [18]. We use a baseline of all 0s and the number of IG iterations is set of 50. Given an output token, the corresponding IG attributions for each input token are a vector of the same dimensionality as the input token embeddings. We convert the vector scores to the a per token scalar score by using the L2 norm reduction, which has been shown to provide similar or better performance to other reduction strategies [40].

Hallucination classifiers are trained with a batch size of 128 for 1, 000 iterations. We use Adam optimizer with a learning rate of  $10^{-4}$  and weight decay of  $10^{-2}$ .

### 4.4 Question Answering Accuracy

Before moving on to hallucination detection, we first analyze the performance of the models in correctly answering the questions i.e., how often the models hallucinate. Table 1 shows the accuracy of different models. The models showed a range of performance across each task. All models performed best on the TriviaQA tasks, and worst on the Birth Place task. Surprisingly, larger models did not always perform better. On the *subject specific tasks* from T-REx, smaller models often performed as well or better than their larger counterparts (6 out of 9 times). On the more *general TriviaQA task*, larger models consistently performed better. On all tasks, FAL-40B significantly outperformed all other models. Further, while larger models on average performed better at general knowledge tasks, variation in performance is more strongly correlated with model type rather than size. For example, while LAM-13B outperformed its smaller variants LAM-7B, and similarly OPT-30B outperformed its smaller variant, both LAM models outperformed both OPT models.

<sup>1</sup>[https://huggingface.co/docs/transformers/main/model\\_doc/open-llama](https://huggingface.co/docs/transformers/main/model_doc/open-llama)

<sup>2</sup>[https://huggingface.co/docs/transformers/v4.19.2/en/model\\_doc/opt](https://huggingface.co/docs/transformers/v4.19.2/en/model_doc/opt)

<sup>3</sup>[https://huggingface.co/docs/transformers/main/model\\_doc/falcon](https://huggingface.co/docs/transformers/main/model_doc/falcon)

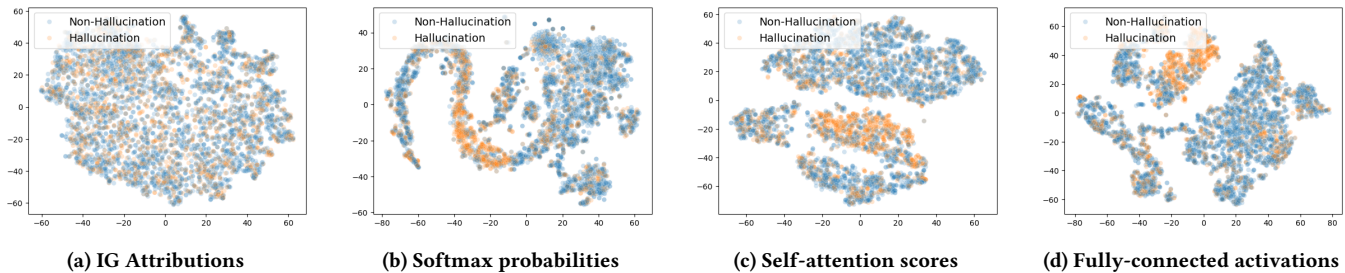
	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.46	0.36	0.67
<b>Capitals</b>	0.33	0.34	0.45
<b>Founders</b>	0.22	0.26	0.33
<b>Birth Place</b>	0.11	0.10	0.06

(a) Large

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.45	0.26	0.54
<b>Capitals</b>	0.44	0.37	0.46
<b>Founders</b>	0.30	0.27	0.30
<b>Birth Place</b>	0.13	0.07	0.07

(b) Small

**Table 1: Accuracy of different models in answering the questions. We consider two model size variants: large (left table) and small (right table). Models tend to perform the best on TriviaQA dataset and the worst on Birth Place dataset. Performance correlates more with the model type than with model size—with FAL performing the best.**



**Figure 3: [FAL-40B on TriviaQA dataset] 2D TSNE projections of the model artifacts for hallucinations and non-hallucinations.**

#### 4.5 Accuracy of the hallucination heuristic

Recall from §3.1 that we consider a model answer  $A$  to be a hallucination if  $R \not\subseteq A$ . To evaluate this heuristic, we randomly sample 100 TriviaQA generations with FAL-40B. For each generation, each of the three authors independently labeled it as (non-) hallucination. Based on the majority vote, **the heuristic was correct in 98/100 cases**. The cases where the heuristic was incorrect are:

**Question:** In Greek mythology, one of the 12 Labours of Hercules was to produce ‘what’ item belonging to Amazonian queen Hippolyte?  
**Ref. Answer:** Magical Girdle  
**Model Answer:** The girdle of Hippolyte was a girdle that was worn by the Amazonian queen Hippolyte. It was one of the 12 Labours of Hercules.  
**Heuristic label:** Hallucination  
**Human label:** Not a hallucination

**Question:** Which British Prime Minister was the 1st Earl of Stockton?  
**Ref. Answer:** Harold Macmillan or Earl of Stockton  
**Model Answer:** Which British Prime Minister was the 1st Earl of Stockton?  
**Heuristic label:** Not a hallucination  
**Human label:** Hallucination

## 5 RESULTS

We first qualitatively investigate the differences in distributions of hallucinations and non-hallucinations and then quantitatively evaluate the performance of hallucination classifiers.

#### 5.1 Qualitative analysis

Recall our hypotheses that hallucinating and non-hallucinating generations differ have different distributions of generation artifacts, namely, Softmax probabilities, IG attributions, self attention and fully-connected activations (§3.2). Examining FAL-40B on the TriviaQA dataset (Figure 3), we find that hallucinated outputs indeed tend to be distributed differently from non-hallucinated outputs when considering Softmax probabilities, self-attention scores and fully-connected activations. The IG attributions however do not show much difference in distributions.

The results when comparing the entropy of distributions, and other datasets and models are mixed (figures omitted due to lack of space). The distributions of input token entropy showed a stronger correlation to dataset type than to model. For topic specific datasets, entropy is on average lower for hallucinated results, while it is slightly higher for hallucinated results on TriviaQA. The entropies of the Softmax outputs are less stable across both datasets and models. While most cases show slightly higher entropy for hallucinated results, the difference is inconsistent. Further, in both the input token and output Softmax cases, the differences, while present, are relatively small. We find similarly small difference in entropy along all generated tokens (see Figure 4). 2-dimensional TSNE plots also show similarly mixed trends.

**While mixed, the results show that in many cases, the reduced distributions of model artifacts (either to 2 dimensions via TSNE or to 1 dimension via entropy) show visually discernible differences between hallucinated and non-hallucinated outputs.** Next we investigate if we can train accurate hallucination detectors by using these artifacts in their original form *i.e.*, without reductions.

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.60	0.57	0.48
<b>Capitals</b>	0.41	0.68	0.43
<b>Founders</b>	0.57	0.44	0.45
<b>Birth Place</b>	0.48	0.45	0.51
<b>Combined</b>	0.43	0.57	0.49

(a) IG attributions

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.71	0.63	0.60
<b>Capitals</b>	0.68	0.69	0.63
<b>Founders</b>	0.61	0.67	0.66
<b>Birth Place</b>	0.66	0.62	0.66
<b>Combined</b>	0.69	0.67	0.62

(b) Softmax probabilities

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.71	0.65	0.71
<b>Capitals</b>	0.72	0.72	0.72
<b>Founders</b>	0.73	0.68	0.71
<b>Birth Place</b>	0.81	0.61	0.81
<b>Combined</b>	0.78	0.71	0.79

(c) Self-attention scores

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.72	0.64	0.72
<b>Capitals</b>	0.73	0.71	0.70
<b>Founders</b>	0.71	0.72	0.73
<b>Birth Place</b>	0.80	0.77	0.76
<b>Combined</b>	0.79	0.73	0.82

(d) Fully-connected activations

Table 2: [Larger model variants] Test AUROC of binary classifiers in detecting hallucinations.

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.62	0.54	0.54
<b>Capitals</b>	0.35	0.34	0.40
<b>Founders</b>	0.40	0.48	0.54
<b>Birth Place</b>	0.42	0.44	0.43
<b>Combined</b>	0.40	0.42	0.53

(a) IG attributions

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.68	0.65	0.64
<b>Capitals</b>	0.65	0.67	0.68
<b>Founders</b>	0.66	0.66	0.72
<b>Birth Place</b>	0.59	0.65	0.64
<b>Combined</b>	0.67	0.67	0.72

(b) Softmax probabilities

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.71	0.66	0.66
<b>Capitals</b>	0.76	0.73	0.68
<b>Founders</b>	0.70	0.72	0.66
<b>Birth Place</b>	0.76	0.64	0.72
<b>Combined</b>	0.75	0.75	0.71

(c) Self-attention scores

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.74	0.70	0.70
<b>Capitals</b>	0.74	0.75	0.73
<b>Founders</b>	0.72	0.71	0.70
<b>Birth Place</b>	0.71	0.71	0.78
<b>Combined</b>	0.77	0.77	0.71

(d) Fully-connected activations

Table 3: [Smaller model variants] Test AUROC of binary classifiers in detecting hallucinations.

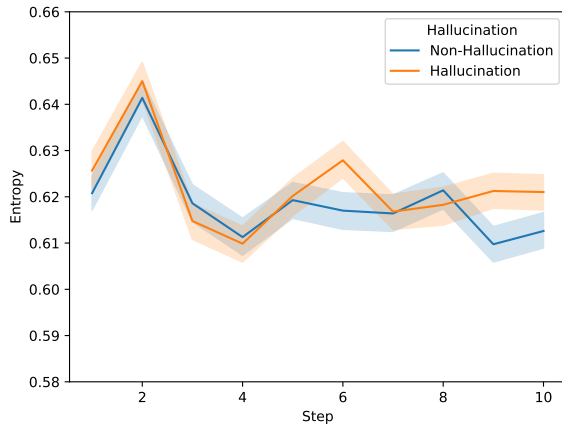
## 5.2 Hallucination Classifiers

Tables 2 and 3 shows the test AUROC when detecting hallucinations. We opt for AUROC instead of binary classification accuracy because of high class imbalance (Table 1). We include the results for binary classification accuracy in Appendix A.

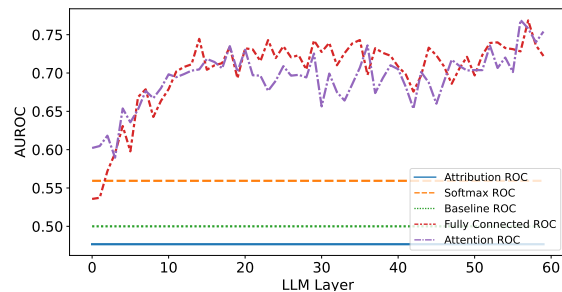
Results show that the IG attribution does slightly better than random chance on the TriviaQA dataset, but no better than random on the subject specific datasets. Softmax consistently does better than random for all tasks. Self-attention scores and fully-connected activations outperform both IG and Softmax. Interestingly, these results hold even though 96% of generated responses start with the newline character, indicating that the classifier is not simply

learning tokens that correlate with hallucination, but that even with the same token, the model internal state between hallucinated and non-hallucinated results differs.

While overall accuracy correlates with model size within each model type, the trend is less consistent for a given model’s ability to identify its own hallucinations. In the case of LAM, the larger variant consistently performed better at identifying hallucination through our classifier. OPT and FAL, on the other hand, showed no consistent correlation to size in their ability to capture hallucinations. In both the large and small model variants, the fully-connected and self attention activation internal states provided the best performance at identifying hallucinations, followed by the model’s Softmax output.



**Figure 4:** [FAL-40B on TriviaQA dataset] Softmax entropy at different generation steps. The difference in entropy between hallucinated and non-hallucinated outputs does not vary much with a change in the generation steps.



**Figure 5:** [FAL-40B on TriviaQA dataset] AUROC of the hallucination detectors using self-attention and fully-connected activations at different layers. The performance is better at later layers but has diminishing returns.

A classifier trained on all four datasets combined (labeled as “Combined” in Tables 2 and 3) tended to perform slightly better than the individual datasets. However, a classifier trained on all model artifacts combined (Softmax, IG, self-attention and fully-connected activations) showed no improvement beyond models trained on each artifact individually (Table 4). There could be several reasons for this: the input dimensionality of this combined classifier is much higher than the individual classifiers and it is architecturally more complex as it consists of both dense and recurrent units. Training this mixed architecture might require special considerations. We leave the detailed analysis to a follow up work.

In summary we note that different model artifacts provide different level of accuracy in detecting hallucinations and **in most cases, self-attention scores and fully-connected activations provide over 0.70 AUROC in detecting hallucinations over a range of datasets and models.**

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.75	0.64	0.65
<b>Capitals</b>	0.69	0.55	0.62
<b>Founders</b>	0.66	0.53	0.63
<b>Birth Place</b>	0.53	0.59	0.68
<b>Combined</b>	0.58	0.58	0.60

**Table 4:** Test AUROC of the classifier trained on all four artifacts at once. The performance is not better than the case when individual artifacts are considered (Tables 2 and 3).

### 5.3 Hyperparameters and baseline comparison

We take a closer look at various hyperparameters and also conduct a performance comparison with an existing method.

**Effect of Transformer layer choice.** The results in §5.2 were based on using the self-attention and fully-connected activations from the last Transformer layer. We also investigate how the performance would change as a result of a change in layer number. Results in Figure 5 show that the performance of our classifier improves with depth of the model layers. Early layers do only slightly better than random chance, while later layers shows significant improvement beyond random chance.

**Effect of classifier hyperparameters.** Given the simplicity of our hallucination detection classifiers, the only hyperparameters available were batch size, learning rate, and weight decay. We hand-selected the defaults in §4.3. We also tried (TriviaQA with FAL-40B) a range of different options for these hyperparameters where we varied batch size from 4 to 256, learning rate from  $10^{-6}$  to  $10^{-2}$  and weight decay from  $10^{-4}$  to  $10^{-1}$ . We found that the smallest batch size and largest learning rate performed worse. But smaller adjustments (batch size 128 vs 256, learning rate  $1 \times 10^{-4}$  v.s.  $2 \times 10^{-4}$ ) made no discernible difference in performance.

We also tested larger models for both our GRU and MLP architectures. On the GRU, we tested from 4 to 12 gated recurrent layers. On our MLP, we tested up to 8 layers, with widths from 32 to 256. In both cases, we found that larger models provided no benefit in either AUROC or accuracy. For example, for the GRU model FAL-40B with TriviaQA, expanding to 12 recurrent layers yielded an AUC of 0.46, slightly worse than the 0.48 reported with 4 layers reported in Table 2. For MLP, expanding to 8 layers yielded 0.72, only slightly better than the 0.71 we reported with the single layer model.

A systematic study of these hyperparameter choices, and understanding the effect of more advanced detector architectures (*e.g.*, Transformers, LSTMs) on the detection performance is a promising future direction.

**Comparison with SelfCheckGPT.** We compare the performance of our method to SelfCheckGPT [24]. Results are shown in Table 5 for larger model variants and Table 6 for smaller model variants. The performance is generally worse than our classifiers. We use two variants of SelfCheckGPT: BERTScore and n-gram. We use 20 generations per input and set the temperature to 0.1. Preliminary analysis in Table 7 shows that setting the temperature to 1.0 as suggested in the paper [24] does not lead to better performance.

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.56	0.58	0.50
<b>Capitals</b>	0.53	0.51	0.56
<b>Founders</b>	0.55	0.49	0.57
<b>Birth Place</b>	0.51	0.54	0.48

(a) BERTScore

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.54	0.53	0.54
<b>Capitals</b>	0.54	0.55	0.54
<b>Founders</b>	0.58	0.56	0.57
<b>Birth Place</b>	0.51	0.49	0.52

(b) n-gram

Table 5: [Larger model variants] AUROC of SelfCheckGPT based on BERTScore and n-gram scores.

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.55	0.56	0.55
<b>Capitals</b>	0.52	0.50	0.49
<b>Founders</b>	0.58	0.58	0.54
<b>Birth Place</b>	0.52	0.51	0.56

(a) BERTScore

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.53	0.58	0.50
<b>Capitals</b>	0.53	0.48	0.49
<b>Founders</b>	0.53	0.43	0.46
<b>Birth Place</b>	0.57	0.58	0.55

(b) n-gram

Table 6: [Smaller model variants] AUROC of SelfCheckGPT based on BERTScore and n-gram scores.

	BERTScore	n-gram
<b>TriviaQA</b>	0.52	0.55
<b>Capitals</b>	0.47	0.50
<b>Founders</b>	0.44	0.51
<b>Birth Place</b>	0.55	0.48

Table 7: [FAL-7B] AUROC of SelfCheckGPT with  $T = 1$ .

## 6 CONCLUSION, DISCUSSION & LIMITATIONS

We address the problematic behavior where large language models hallucinate incorrect facts. By extracting different generation artifacts such as Integrated Gradients feature attribution scores and self-attention scores at various Transformer layers we build simple classifiers for detecting hallucinations. The detection can already be performed at the start of the response generation, that is, factual hallucinations can often be detected even before they occur. We show that the ability to identify hallucinations persists across different topics like Capitals, Founders and Birth Place; and in a broader trivia knowledge context. Attaching classifiers of this type to deployed LLMs can serve as an effective method of flagging potentially incorrect information before it reaches the end-users or downstream applications.

Surprisingly, the classifiers are able to detect hallucinations even when the first generated token is a seemingly “uninformative” token like a formatting character. With the exception of Falcon models on the Birth Place dataset, the newline character is the first token in 99.5% of generated responses, with the remaining 0.5% being a punctuation. In the Falcon / Birth Place case, the newline character starts 82% of responses, while the word “where” starts the other 18%. Two factors likely contribute to this surprising effectiveness of such “uninformative” tokens. 1) while the first token `\n` is a single character, the associated artifacts like hidden states and softmax

probabilities contain much more information as these are vectors with thousands of dimensions. 2) We are experimenting with autoregressive models. At the first generation location, the model has ingested all the external information (that is, the input) and the upcoming generation only depends on the internals of the model itself (minus the effect of the generation strategy which in our case is the most likely next token). Nonetheless, this finding merits further investigation and is a promising avenue for future work.

We only considered factual hallucinations in this work. It would be interesting to study if the proposed method also extends to other forms of hallucinations [13, 14]. We also considered a rather simple setup where the model output is supposed to contain a single fact only. Extending our method to probe multiple hallucinations (e.g., multiple facts in a biography like place of birth, date of birth, alma mater) in a single generation is also an important follow up direction.

Our method cannot be used with “blackbox” models that do not reveal model internals and gradients to the users. Text-only methods like SelfCheckGPT are well-suited for such situations but compensate for the lack of internal access by querying the model several times. In our case, all except one artifact can be computed “for free” during the forward pass. Nonetheless, our method can still be deployed by the model providers internally. Experimenting with more elaborate retrieval settings (e.g., those using retrieval augmented generation [19], instruction tuned models [29] and in-context examples [39]), a wider range of datasets and model types is also a promising avenue for future work.

Finally, the goal of this paper was not to extract the maximum possible detection performance, but to test if certain artifacts can be promising in detecting hallucinations. For this reason, we use relatively simple detection architectures. We leave a more in-depth analysis of detection architectures to a future study.

The code repository for the paper is available at: <https://github.com/amazon-science/llm-hallucinations-factual-qa>.

## REFERENCES

- [1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 11, 681–699. doi: 10.1162/tacl\_a\_00667.
- [2] Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.
- [3] Jon Christian. 2018. Why Is Google Translate Spitting Out Sinister Religious Prophecies? <https://www.vice.com/en/article/j5npeg/why-is-google-translation-spitting-out-sinister-religious-prophecies>. Accessed: 2023-08-16. (2018).
- [4] 2023. Confirmed: the new Bing runs on OpenAI's GPT-4. [https://blogs.bing.com/search/march\\_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4](https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4). Accessed: 2023-10-11. (2023).
- [5] David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: model internal workings alone do well, sentence similarity Even better. In *ACL* (July 2023).
- [6] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [7] Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: a large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, (May 2018). <https://aclanthology.org/L18-1544>.
- [8] Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 8756–8769. doi: 10.18653/v1/2022.emnlp-main.599.
- [9] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8, 539–555.
- [10] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: an overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
- [11] Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: a comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, (May 2023), 1059–1075. <https://aclanthology.org/2023.eacl-main.75>.
- [12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51, 5, 1–42.
- [13] Lei Huang et al. 2023. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- [14] Ziwei Ji et al. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55, 12, 1–38.
- [15] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, (July 2017), 1601–1611. doi: 10.18653/v1/P17-1147.
- [16] Saurav Kadavath et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- [17] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, (Eds.) Association for Computational Linguistics, Toronto, Canada, (July 2023), 5591–5606. doi: 10.18653/v1/2023.acl-long.307.
- [18] Narine Kokhlikyan et al. 2020. Captum: a unified and generic model interpretability library for pytorch. (2020). arXiv: 2009.07896 [cs.LG].
- [19] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- [20] Percy Liang et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- [21] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, (Eds.) Association for Computational Linguistics, Dublin, Ireland, (May 2022), 3214–3252. doi: 10.18653/v1/2022.acl-long.229.
- [22] Zachary C Lipton. 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue*, 16, 3, 31–57.
- [23] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [24] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali, (Eds.) Association for Computational Linguistics, Singapore, (Dec. 2023), 9004–9017. doi: 10.18653/v1/2023.emnlp-main.557.
- [25] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35, 17359–17372.
- [26] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209.
- [27] 2023. Need a Last Minute Mother's Day Gift? AI Is Here to Help. <https://about.you.com/need-a-last-minute-mothers-day-gift-ai-is-here-to-help-d363b17e76b4/>. Accessed: 2023-10-11. (2023).
- [28] OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [29] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- [30] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: a benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.
- [31] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 2463–2473. doi: 10.18653/v1/D19-1250.
- [32] Sundar Pichai. 2023. An important next step on our AI journey. <https://blog.google/technology/ai/bard-google-ai-search-updates/>. Accessed: 2023-10-11. (2023).
- [33] Nathalie Rauschmayr et al. 2021. Amazon sagemaker debugger: a system for real-time insights into machine learning model training. In *MLSys 2021*. <https://www.amazon.science/publications/amazon-sagemaker-debugger-a-system-for-real-time-insights-into-machine-learning-model-training>.
- [34] Stephen Roller et al. 2020. Open-domain conversational agents: current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.
- [35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [36] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: a strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3, 6, 7.
- [37] Timm Teubner, Christoph M Flath, Christof Weinhart, Wil van der Aalst, and Oliver Hinz. 2023. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65, 2, 95–101.
- [38] Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, (Aug. 2021), 1126–1140. doi: 10.18653/v1/2021.acl-long.91.
- [39] Qinyuan Wu et al. 2024. Towards reliable latent knowledge estimation in llms: in-context learning vs. prompting based factual knowledge extraction. *arXiv preprint arXiv:2404.12957*.
- [40] Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cédric Archambeau, Sanjiv Das, and Krishnamurthy Thakur. 2021. On the lack of robust interpretability of neural text classifiers. *arXiv preprint arXiv:2106.04631*.
- [41] Chrysoula Zerva et al. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, 961–972.
- [42] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

## A ADDITIONAL RESULTS

Recall that due to the imbalance between hallucinations vs. non-hallucinations (Section 4.4), we reported AUROC in the main experiments (Section 5.2). Tables 8 and 9 show the binary classification accuracy for larger and smaller model variants.

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.64	0.67	0.66
<b>Capitals</b>	0.68	0.65	0.63
<b>Founders</b>	0.70	0.69	0.68
<b>Birth Place</b>	0.85	0.87	0.89

(a) IG attributions

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.65	0.63	0.68
<b>Capitals</b>	0.70	0.71	0.63
<b>Founders</b>	0.73	0.68	0.67
<b>Birth Place</b>	0.86	0.91	0.92

(b) Softmax probabilities

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.69	0.65	0.70
<b>Capitals</b>	0.70	0.71	0.66
<b>Founders</b>	0.72	0.74	0.68
<b>Birth Place</b>	0.89	0.91	0.93

(c) Self-attention scores

	LAM-13B	OPT-30B	FAL-40B
<b>TriviaQA</b>	0.67	0.65	0.69
<b>Capitals</b>	0.69	0.70	0.65
<b>Founders</b>	0.73	0.74	0.68
<b>Birth Place</b>	0.88	0.89	0.94

(d) Fully-connected activations

Table 8: [Larger model variants] Test accuracy of binary classifiers in detecting hallucinations.

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.63	0.68	0.63
<b>Capitals</b>	0.70	0.64	0.67
<b>Founders</b>	0.69	0.70	0.68
<b>Birth Place</b>	0.83	0.88	0.89

(a) IG attributions

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.66	0.63	0.66
<b>Capitals</b>	0.67	0.68	0.64
<b>Founders</b>	0.71	0.63	0.68
<b>Birth Place</b>	0.87	0.90	0.89

(b) Softmax probabilities

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.65	0.62	0.72
<b>Capitals</b>	0.70	0.72	0.65
<b>Founders</b>	0.69	0.71	0.69
<b>Birth Place</b>	0.91	0.90	0.91

(c) Self-attention scores

	LAM-7B	OPT-6.7B	FAL-7B
<b>TriviaQA</b>	0.66	0.68	0.66
<b>Capitals</b>	0.70	0.69	0.65
<b>Founders</b>	0.71	0.72	0.70
<b>Birth Place</b>	0.87	0.88	0.93

(d) Fully-connected activations

Table 9: [Smaller model variants] Test accuracy of binary classifiers in detecting hallucinations.