# Hallucinations in Large Language Models

**Nicolò Brunello***
DEIB, Politecnico di Milano
Via Golgi 42, 20133, Milano (MI), Italy
nicolo.brunello@polimi.it

Wednesday 1st April, 2026

**Abstract**

The widespread adoption of Large Language Models (LLMs) has intensified concerns about their trustworthiness, particularly due to their tendency to generate hallucinations: outputs that are fluent and persuasive yet factually incorrect, unfaithful to the input, or logically inconsistent. This chapter surveys the phenomenon of hallucination in LLMs through a structured analysis of its definitions, causes, detection methods, mitigation strategies, and future research directions.

We begin by showing that hallucination is not a uniquely defined concept in the literature, and we review the main taxonomies used to characterize it, including distinctions between intrinsic and extrinsic hallucinations, factuality and faithfulness errors, and instruction-, context-, and logic-related inconsistencies. We then examine the major sources of hallucinations across three dimensions: training data, training and inference procedures, and model architecture. In particular, we discuss how data quality issues, knowledge limitations, alignment effects, decoding randomness, and attention-related failures may all contribute to unreliable model behavior.

The chapter further presents a review of hallucination detection techniques, from lexical overlap metrics and classifier-based approaches to uncertainty estimation, multi-agent evaluation, and retrieval-based verification against external knowledge sources. Finally, we discuss emerging directions centered on interpretability and mechanistic analysis of Transformer models, arguing that a deeper understanding of internal model computations may play a crucial role in reducing hallucinations and improving reliability. Taken together, the chapter frames hallucinations as a fundamental and multi-faceted limitation of contemporary LLMs, and highlights the need for rigorous definitions, robust evaluation protocols, and more transparent model development practices.

**Keywords:** one • two • three

## 1 Introduction

As artificial intelligence (AI) continues to revolutionize various industries, one of its most intriguing and impactful challenges is the trustworthiness of these models. All machine learning algorithms come with an intrinsic uncertainty, tightly bound to their stochastic inductive nature to generalize from a finite set of data. This is not a problem in many applications, nor a limitation of their capabilities, still, in many industries they struggle to find a fertile soil to ground their usage and grow the number of real-world applications effectively relying on machine learning algorithms.

In this scenario, perfect examples are the well-known Large Language Models (LLMs). These advanced systems are capable of generating human-like text, providing valuable insights, and assisting in tasks from content

*Corresponding author

creation to technical support. However, with their growing influence comes a critical question: Can we trust these models?

A key challenge with LLMs is their tendency to produce what has been termed "hallucinations"—responses that may sound coherent but are factually incorrect or entirely fabricated. These hallucinations can have far-reaching consequences, from spreading misinformation to undermining the credibility of organizations that rely on these models.

In this chapter, we explore the trustworthiness of AI by delving into the mechanics of LLM hallucinations, why they occur, and how they impact real-world applications. We'll examine the current state of research and development focused on mitigating these issues, as well as practical strategies for organizations to deploy AI responsibly. Through this journey, we aim to provide a comprehensive understanding of the promise and limitations of LLMs, shedding light on what it means to trust AI in an era of rapid technological advancement.

As we navigate the complex terrain of AI trustworthiness, we must ask ourselves: How can we harness the power of these models while ensuring the integrity of the information they produce?

## 2 Definition

The term *hallucination* was first used back in the days by the Neural Machine Translation community (Raunak et al., 2021; Müller et al., 2020) referring to erroneous output that is fluent in the target language but, at the same time, decoupled from the source sequence. But, why did they choose exactly the term "hallucination"? It calls back to human psychosis-related disorders like schizophrenia and dementia, whose hallmark is represented by false perceptions of sensory experiences. Indeed, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines hallucinations as "perception-like experiences that occur without an external stimulus" and which "are vivid and clear, with a full force and impact of normal perceptions, (though) not under voluntary control" (American Psychiatric Association, 2022). By transposing this definition in the text LLMs scenario, a hallucination is commonly referred to as an *undesired phenomenon of Natural Language Generative models generating unfaithful or nonsensical text* (Ji et al., 2023); so this interpretation, though it seems plausible, is only suitable to grasp the generic meaning of what exactly is a hallucination and not how could be precisely identified in real cases. There is a well-deserved discrepancy between theoretical dissertations about LLMs and real examples due to their breakneck growth in potential: LLMs are now so flexible that their applications are no longer restricted to a single task, instead, they leverage the power of these models to manipulate text nearly in every domain and every task, even by calling functions and changing an application workflow. Such versatility, combined with the stochastic nature of LLMS, leads to undesired outputs that may be largely different from each other, making it hard to reliably identify hallucinated content. For this reason, we now provide some taxonomies and additional definitions to better analyze the hallucination phenomenon.

### 2.1 Taxonomies and examples

We now delve into the different definitions and examples by reviewing the current state-of-the-art literature, but before addressing a comprehensive review, we need to state some basic components and definitions of LLMs-based applications.

Many LLMs are nowadays used along with a retrieval engine which enhances the LLMs knowledge with some specific documents and information. We can identify:

- **Query** A user statement, that may or may not be in the form of a question, which contains what a generic user is expected from the LLM. Some examples of queries are: "What is the capital of Rome?" or "List all the countries in the European Union".
- **Context** Pieces of documents returned by a retrieval engine that are related to a query and should be used by the LLMs as ground truth to answer the query.

- **Response** The final answer of the LLM to the query

Given these basic components, we can now state some definitions useful to fully explain each category of hallucinations.

**Definition 2.1** (Hallucination)**.** Generation of text which is nonsensical or unfaithful with respect to the provided source input (Ji et al., 2023; Maynez et al., 2020; Filippova, 2020)

**Definition 2.2** (Faithfulness)**.** Quality of a piece of text staying consistent and truthful with respect to some provided sources.

**Definition 2.3** (Factuality)**.** Quality of a piece of text being actual or based on fact. Depending on what serves as the “fact”, "factuality" and "faithfulness" may or may not be the same.

**Definition 2.4** (World knowledge)**.** What is commonly considered to be true, despite source input text.

Already from these preliminary definitions, it is easy to see that the task of defining hallucinations is all but trivial. For example Maynez et al. (2020) differentiate "factuality" from "faithfulness" by defining the “fact” to be what is true according to the world knowledge, i.e. what is true in reality, but, is it always possible to say whether a fact is true or false? Unfortunately, if the answer was positive there would not be complications like fake news or misinformation in our everyday life, but is out of the scope of this work to end in ethical or philosophical dissertations.

Let’s start with a simple yet effective consideration:

*Remark.* The knowledge base inside a Large Language Model contains at most the same information that is contained in the pre-training dataset.

The pre-training dataset is usually composed of a huge amount of text scraped from the Internet and verifying the quality of each chunk of text is not a feasible task. As we will see in Section 3 this is the first flaw of LLMs and, consequently, the major source of hallucinations. Moreover, at the time this work was written, is it not clear how to assess whether a given information has been "learned" by the Transformers models or not, i.e. there is no way of precisely mapping the internal knowledge base of an LLM if not by prompting questions and manually checking answers. This leads to the uncontrollability of the information an LLM has acquired during the pre-training and the necessity to restrict the controllable knowledge to a source text used as ground truth.

For these reasons, the very first distinction among hallucination types is (Dziri et al., 2021; Maynez et al., 2020; Ji et al., 2023):

- **Intrinsic** Given a *query* and a source *context*, the *response* from the LLM straightforwardly contraddicts the source text. For example, if the query is *"What color could be an apple?"*, the source context is the sentence *"The apples could be green, yellow and red."* and the response of the LLM is *"The apples could only be blue"*. This is an intrinsic hallucination since the response is directly in conflict with the context.
- **Extrinsic** Given a *query* and a source *context*, the *response* could not be verified by the source text. An example is, given the same query and context as before, the response is *"All apples are delicious!"*. This is an extrinsic hallucination because, if we only consider the context, the response is not explicitly stated, nor deducible.

This first categorization was as simple as effective since let researchers start exploring causes and mitigations given a context, which is a much simpler task to tackle with respect to comparing the outputs of LLMs with their entire knowledge base. Nonetheless, this categorization is missing an important use case, which is nowadays commonly adopted in the majority of LLMs applications, that is *instruction following*: assessing if the LLMs have correctly followed the instructions in the prompt is totally different task with respect to checking the correctness of the content, and these two jobs are not mutually exclusive.

To overcome these limitations, more recent works attempted to provide a more comprehensive and detailed classification of hallucinations. Huang et al. (2023) proposed a slightly more granular categorization:

- **Factuality Hallucination** The output is inconsistent with real-world facts (i.e. *world knowledge*) or potentially misleading. This first definition resembles the one given before, as a matter of fact, it is subsequently divided in

  - **Factual Inconsistency** The *response* is in direct conflict with the *world knowledge*. For example, if the query was *"Tell me about the first person to land on the Moon."* and the response was *"Yuri Gagarin was the first person to land on the Moon"* it is in direct conflict with a well-known fact that Neil Armstrong was the first person to land on the Moon (Yuri Gagarin was the first person in space instead). Note the reference to the intrinsic hallucinations of the previous categorizations, with the difference that here the response is directly compared with the world knowledge. The flaw of taking world knowledge into account is that it is often disputed what is true for everybody, for example in this example some people do not even believe that humans have ever been to the Moon, of course, these are corner cases so they are not really taking into consideration in this review, but they are worth mentioning to understand the complexity of deciding what is an hallucinations and what is not.
  - **Factual Fabrication** The *response* could not be verified, so there is no evidence supporting or contradicting it. For example, if the query was *"Tell me about the historical origins of unicorns"* and the response was *"Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC"* we can't actually demonstrate the presence of such creatures in a mythological ancient society, so there is no way of actually asserting or contradicting the response. Note the resemblance to the extrinsic hallucinations.

- **Faithfulness Hallucination** Faithfulness has a broader meaning with respect to factuality, it also embraces concepts like instruction following or consistency with the context. Faithfulness hallucinations are all types of hallucination coming from the interaction with a user or a context and could be further divided into three categories.

  - **Instruction inconsistency** Given some *query* prompt containing some instructions, the *response* from the model does not follow those instruction. For example, if the query was *Translate the following English question into Spanish: "What is the capital of France?* and the response was *"The capital of France is Paris"* the model clearly did not understand the task correctly.
  - **Context inconsistency** Given some *query* prompt containing some *context* information, the *response* directly contraddicts those information. An example, quite common in RAG domain, is given by a query *"How tall is the Tour Eiffel?"*, the context is *"The tower is 330 metres (1,083 ft) tall, about the same height as an 81-storey building, and the tallest structure in Paris"* and the response is *"The Tour Eiffel is 1,083 metres tall"*. In this case, even though the final meaning of the sentence has not been completely changed, there is a crucial inaccuracy in interpreting feet as meters, and so, having a final height extremely wrong.
  - **Logical inconsistency** Given some *query* prompt containing some *context* information, the *response* contains some steps of reasoning with some mistakes among them. For example, given the query *"Nick has 10 apples. He sold 4 apples. He divided the remaining apples into two halves and ate one of these halves. Then, he bought 2 more apples, how many apples have now Nick? Please, answer by producing a step-by-step reasoning."* and the response of the LLM is *"At first Nick had 10 apples, he sold 4 apples so 10 - 4 = 6. Now Nick has 6 apples and he ate half of them, so 6 - (6/2) = 6 - 3 = 3. Now Nick has 6 apples and bought 2 more apples so 6 + 2 = 8. So Nick now has 8 apples."*. From this response, it is easy to spot incoherence with the generated response itself, due to mistakenly writing the intermediate total number of apples, after the correct steps of subtraction and division. These kinds of errors are commonly encountered in RAG scenarios where very long reasoning, especially with many numbers, takes place to produce structured outputs.

The categorization proposed so far focuses on incongruences inside the content, though there could be other examples (Zhang et al., 2023), we would like to focus on other types of categorization. For instance, Wang et al. (2023) divide the different types of hallucinations based on what are the flawed components that may have produced the hallucination. They firstly separated the "Factuality" issue from the "Hallucination" issue by dividing the hallucinations into the following types (note that for this categorization we are not providing specific examples since the original ones belong to some outdated versions of current state of the art LLMs).

- **Model-level causes** Issues in the LLMs itself could lead to hallucinations in the output.
  - **Domain knowledge deficit** The model lacks specific knowledge from its pre-training dataset (or eventually the subsequent fine-tuning datasets). The frequency of these hallucinations is decreasing due the the always greater amount of data that is nowadays fed to the LLMs pre-training. Nonetheless, asking for any specific information about some private company knowledge base, for example, will inevitably lead to some wrong answer or, if the model was well aligned, to some pre-formatted answer.
  - **Outdated information** Every internal knowledge base of a LLMs is composed by data up to the date of creation of the dataset, so newer events, or update to existing information, could not be known by the model unless another training is performed.
  - **Reasoning errors** Wrong deduction made by the model while generating an answer. This is related to the previous Logical Inconsistencies highlighted in the previous categorization
- **Retrieval-level causes** Issues related to the gathered context returned by the retrieval engine.
  - **Distracted by the retrieval information** The context may confuse the LLM that could focus more on the context and forgetting the initial query. This may happen when context is too long, or could also be used by a malicious user who is exploiting this property of LLM and induce a wrong generation (Toyer et al., 2024).
  - **Misuderstood the retrieval information** The LLMs wrongly interprets some data in the context retrieved. This could be both a fault of the context, which may be unstructured and not organized, but also fault of the LLM, that could be not powerful enough to understand a specific new topic.
  - **Fail to address misinformation in retrieved context** This is a very specific category aiming at highlighting all those wrong output were the context is structured in some particular way, such as the LLMs should pay attention only to a subpart of it, and leave all the other (e.g. a series of news, one more recent of the other, where some news is updating the information of the previous one).
- **Inference-level causes** Hallucinated outputs due to the inner nature of LLMs of producing text one token at a time.
  - **Snowballing** Phenomenon in which the LLMs, while producing a response, pick as next token a suboptimal token, consequently compromising the entire sequence by amplifying the mistake in the semantic meaning, for the sake of fluency (or other secondary properties).
  - **Exposure bias** Refers to the tendency of generating output similar to the majority of text seen during pre-training. In this category, a perfect representative is gender bias, where the LLMs, if not explicitly specified, tends to generate content regarding certain stereotypes.

Many other categorizations are present in literature (OpenAI, 2023; Kaddour et al., 2023), but listing all of them is out of the scope of this work since new use-cases and new definitions come up every day. However, with this overview, the author would like to point out some key takeaways:

*Remark.* It is currently missing a *quantitaive* definition of hallucination, that is to say, some clear, objective, measurable metric that indisputably characterizes hallucinations.

*Remark.* Without a quantitative definition, it is hard to operationalize the detection of hallucination and, consequently, the trustworthiness of Large Language Models.

Given all these definitions, occasionally very different from one another, is it even worth asking ourselves "Is it possible to make the LLMs more trustworthy?" How can we delve into the complex heuristics and ambiguities of human language to extrapolate specific patterns characterizing such a nebulous phenomenon? The peculiar characteristic of the LLMs, thus of the hallucinations, is the common *social adoption*. Thanks to the advent of ChatGPT (OpenAI, 2022), LLMs became an easily accessible resource that everyone could benefit from. Indeed, a chat-based interaction environment is familiar to most potential ChatGPT users, who started using the tool in their everyday lives, thus, leading to widespread adoption and popularity. A huge community of users and practitioners has entered the game, in addition to all the researchers who were actively studying LLMs, and suddenly the term "hallucination" became a word used to refer various different phenomenon, often affected by subjectivity of the user (i.e. "What for me is an hallucination may not be the same for you"). This procedure notably increased the speed of in research but it is, nonetheless, worth noticing that it is non-standard for a new technology to be adopted by public users without an ontological analysis and definition of one of its main pitfalls. Luckily, this lack of precise definitions is compensated by a thrilling and active research environment, especially in the field of explainability. In the next sections, we will present an overview of the causes identified so far, promising directions, and detection and mitigation methods.

# 3 Hallucination causes and mitigations

In the previous section, we highlighted the intrinsic nature of hallucinations as vague and not well-defined. Such characteristics feed through the intricate task of finding possible causes of hallucinations. In science, when a phenomenon is observed, an investigative mind is naturally prone to find the sources of such phenomenon, leading to advances in theoretical ontological definitions and more practical applications. Declined in the case of hallucinations, which are commonly associated with an undesired behavior of the model, the main driving factor in finding the source(s) is to find possible mitigation techniques, in order to reduce the occurrences of hallucinated content. Despite there is not a precise and generalized definition, many different causes have been identified as "roots" of hallucinations, by restricting the general meaning of "hallucinated content" to specific cases.

We are now presenting the causes, and corresponding mitigations, of hallucinations identified by different studies, divided according to the component affected: *Training data*, *Training and Inference procedure* and *Model architecture*.

## 3.1 Training Data

The pre-training dataset is the core component that shapes the Language Model; every reasoning faculty, task-solving ability and knowledge acquired comes from the pre-training dataset. Common small-sized open-source models like Mistral (Jiang et al., 2023), LLama 3 (Dubey et al., 2024), Olmo (Groeneveld et al., 2024), Gemma 2 (Rivière et al., 2024), have a pre-training dataset size that approximately ranges from 2 to 15 trillions of tokens of text. Such a huge amount of text could not be found from a single and controlled source, that is desirable to have high-quality data, but the majority of it must be scraped from the web. To better give an idea of the size of a pertaining dataset, the whole Wikipedia dump is, at the time of this work, about 24.09 GB of pure text data composed of 4.7 billion words. For the sake of the example, to translate the number of words in the number of tokens we will just use the commonly accepted approximation $words = 0.75 \times tokens$, which results in a total number of tokens of approximately 6.26 billion, less than $0.3\%$ of a small-sized pre-training dataset. This is the reason why the pre-training dataset is often gathered from scraping web data; a straightforward consequence of this approach is the quality loss of data. The reader can easily imagine conversations taken from online forums and social media being, on average, quite far from the ideal text suitable for learning both a language or just

general knowledge: everyone can make up and spread false information, give negatively biased opinions, and use tones not worth of imitation, not to mention grammatically wrong messages or language obscenities.

**Scaling issues** Controlling and filtering the pre-training dataset is a hard, yet essential task to produce more accurate and trustworthy LLMs. Due to its extremely large size, for example, it is not feasible to precisely map every single occurrence of a topic and the different perspectives used to present it. In this scenario, great works studied the scaling of a model and tried to quantify the size of a pre-training dataset in relation to its number of parameters (Hoffmann et al., 2022), justifying the very big size of the dataset (the more parameters, the more tokens needed). More recent studies, further enhance this relation by demonstrating that less data could lead to state-of-the-art results if their quality is drastically improved (Gunasekar et al., 2023).

**Examples of data flaws** Survey works like Ji et al. (2023); Huang et al. (2023) highlight the great literature concerning the relation of data quality and LLMs performance, showing some peculiar cases. For example, since the dataset is taken from text written by any Internet user, there is no guarantee concerning its actuality. Without considering the edge case of an intentionally not factual statement, let's consider the so-called ***Imitative falsehoods*** (Lin et al., 2022), common misconceptions or unprecise statements that are actually not true. A simple example is *"What percentage of the brain does a human typically use?"* and the answer of the model is *"A human typically uses about 20 percent of his or her brain."*, which is an answer based on the fact that the model saw this concept many times during training since it is a common belief without real scientific grounding.
Another similarly interesting case is about duplicated statements (Hernandez et al., 2022). LLMs, like many other types of massive Neural Networks, due to their intrinsic stochastic nature are prone to overfit the data which, translated in the context of language modeling, ends in the generation of sequences matching exactly some memorized chunk of training data. This phenomenon is called ***duplication bias*** and is shown to be more likely while the model size is increasing (Lee et al., 2022).
Another type of pervasive bias in data is ***social bias***. These range from biases related to gender, religion, and race to those involving occupation and other subjects (Kotek et al., 2023). These biases mirror the distribution of text used in pre-training, which in turn reflects societal patterns—hence the term "social bias". An example of gender bias is the model assuming a nurse is female, even when the gender is not explicitly specified.
The last main type of data flaw is the limitation of knowledge contained in the pre-training data, commonly referred to as ***knowledge boundary***. Although the dataset is vast, its knowledge is far from uniformly distributed; it leans heavily toward certain popular topics, leaving the more niche or "long-tail" areas relatively underrepresented.

**Mitigation** To provide a comprehensive understanding of the techniques used while filtering a pre-training dataset, thus mitigating the phenomena shown before, we are now reviewing the operations made by some open-source models and relating them with the rest of the literature. Soldaini et al. (2024) precisely shows the construction of Dolma (Data for Open Language Models' Appetite), a dataset composed of 3T tokens english corpus for pre-training language models. The initial step is data acquisition, primarily sourced from the Common Crawl repository[1], which is recognized as the largest publicly available collection of online text, immediately followed by an initial filtering of repetitions by leveraging CCNet (Wenzek et al., 2020). Subsequently, a *cleaning step* extracts only pure text from HTML documents, discarding all non-prose content like tags, headers, and decoration text, using quality filters from (Rae et al., 2021) (each filter combination has been validated on a common benchmark). In the remaining documents, *content filters* are utilized to minimize toxic and harmful content; these filters include trained classifiers that have been calibrated on the data distribution through ablation studies. Additionally, an important filter is the *PII filter* (Personally Identifiable Information), which aims to eliminate any references to individuals to prevent the model from inadvertently revealing them during inference. Given the dataset's large size, the authors opted for regular expression matching rather than model-based techniques for PII removal. Finally, the step of *paragraph-level deduplication* removes copies of the same

[1] https://commoncrawl.org/

paragraphs across all documents.
In constructing Llama 3 (Dubey et al., 2024), similar steps have been applied to training data:

- *Parsing HTML* with custom parsers to remove boilerplate characters, menu, navigation headers, etc., and keep only significant text. Particular attention had been spent on math texts, often appearing as a pre-rendered image.
- *Deduplication* slightly more aggressive deduplication than before, removing even lines appearing more than 6 times in the same batch of data.
- *Heuristic filtering* similar to the previous ones (Rae et al., 2021), plus a filter that leverages Kullback-Leibler divergence on token distributions to remove similar documents.
- *Model-based quality filters* using trained classifiers and scoring models, empirically validated to calibrate the optimal combination.

.

With this overview of techniques for building a pre-training dataset, we have highlighted state-of-the-art methods to reduce occurrences of hallucinations and other harmful content. The core idea is to ensure high-quality data, which we define as data that is accurate, relevant, and representative of diverse perspectives while being free from toxic or misleading elements. By prioritizing quality in the dataset, we can enhance the model's reliability and safety, ultimately leading to more trustworthy outcomes in its applications.

### 3.2 Training and Inference procedure

We will now explore other types of possible causes of hallucinations, specifically from training procedures and inference methods. Before actually delving into the explanations, it is useful to recall the different steps in training and the different inference methods used to build Large Language Models. Following the pre-training process described in the previous section, an *alignment* step is typically applied (Ouyang et al., 2022). After completing pre-training, an LLM is not yet suited for use as an assistant, where it must follow user instructions and perform tasks under specific conditions. To address this limitation, the alignment step is introduced to "align" the model's outputs with user expectations. This alignment is achieved through Reinforcement Learning based on human-annotated preferences. The LLM has now gone through the main steps needed to build the core foundation model, so it can be used to do inference by generating tokens given some user input. Given a sequence, the next token is sampled from a probability distribution over all the vocabulary space, produced by passing the last hidden layer of the Transformer to a Language Modeling head (Neural Network + Softmax activation). The action of sampling the next token is called "decoding", and involves many hyperparameters to tune properly to avoid degeneration of output sequences.

**Alignment** Aligning the language model to human preferences is a necessary, yet dangerous task. During pre-training, the model develops a set of underlying abilities to predict the next token, and in alignment, it should leverage these abilities to generate coherent and effective output text. However, because these abilities are not directly observable, alignment can sometimes push the model beyond its natural capabilities, effectively "forcing" it to acquire certain skills. In the worst-case scenario, this can lead to overfitting, where instead of genuinely learning these abilities, the model begins to replicate specific text patterns, which can undermine its internal knowledge and overall capabilities. This phenomenon is referred to as *Capability Misalignment*. Additionally, the model has been shown to develop internal confidence in its generated content (Azaria and Mitchell, 2023). However, if the alignment step is too strict, the model may prioritize "satisfying" user preferences over producing truthful text, a behavior known as *sycophancy* (Cho et al., 2023).

**Decoding** The decoding phase, while not a training step, is crucial for harnessing the full potential of an LLM. During decoding, the model samples the next token from a probability distribution over the entire vocabulary space, introducing an inherent element of randomness. Typically, LLMs in production avoid using greedy decoding, i.e. consistently selecting the highest probability token, which can lead to repetitive and predictable outputs. Instead, various sampling techniques—such as top-k sampling, top-p (nucleus) sampling, and temperature scaling—are employed to balance coherence and diversity in the generated text. These methods allow the model to generate both contextually relevant and varied responses, enhancing the naturalness and engagement of its interactions with users, and avoiding the undesired phenomenon of *likelihood traps* (Zhang et al., 2020)—the counter-intuitive observation that high likelihood sequences are often surprisingly low quality. However, introducing randomness comes at a cost as increasing diversity in the output enhances the chances of producing nonfactual or unfaithful text (Dziri et al., 2021).

**Mitigation** Enhancing alignment in LLMs can be significantly improved through high-quality annotation. Accurate, context-aware annotations offer several benefits: they help models better understand nuanced user instructions, reduce tendencies toward generating harmful or biased content, and encourage more truthful and relevant responses. Clear and comprehensive annotations also provide a more robust foundation for models to generalize to new prompts, creating more consistent user experiences.
Although not a training step, the decoding phase is essential for maximizing an LLM's performance. To avoid repetitive or overly deterministic responses, LLMs often rely on sophisticated sampling techniques, such as adjusting temperature, top-p (nucleus sampling), and top-k, rather than using simple greedy decoding.

Proper calibration of these decoding parameters is key to balancing coherence with creativity and achieving responses that are relevant, diverse, and aligned with user intent:

- *Temperature*: The temperature parameter adjusts the "sharpness" or diversity of the model's responses by scaling the logits before sampling. A lower temperature (close to 0) makes the model more deterministic, often leading to focused and predictable responses by favoring high probability tokens. However, overly low temperatures can result in repetitive output, especially for complex tasks requiring creativity. Higher temperatures (e.g., 0.7 to 1.2) introduce more variety by flattening the probability distribution, which allows the model to explore less likely tokens. Finding the right temperature depends on the task: for factual responses, lower temperatures are often preferable, whereas creative or open-ended tasks benefit from moderate to higher values.

- *Top-K Sampling*: In top-k sampling, the model limits its choice to the top-k most likely tokens, rather than considering the entire vocabulary. This restricts randomness to a manageable scope, making responses coherent while retaining some diversity. For instance, setting k between 5 and 50 allows the model to produce coherent and contextually appropriate responses without straying into unlikely or irrelevant tokens. Lower k-values make responses more predictable, while higher k-values introduce greater diversity. Selecting an optimal k-value is task-dependent; lower values are suitable for deterministic, factual responses, whereas higher values can be applied to conversational or creative contexts.

- *Top-P (Nucleus) Sampling*: In contrast to top-k, top-p sampling dynamically selects a subset of tokens based on cumulative probability. The model selects from the smallest group of tokens whose probabilities sum to at least the chosen p-value (e.g., 0.9). This adaptive approach allows for variable token counts based on the context of each prediction, balancing flexibility and coherence. Lower p-values (such as 0.8) reduce diversity and encourage more straightforward responses, ideal for generating factual content. Higher values (e.g., 0.9 or 0.95) allow for greater variance in output, suitable for creative applications or generating natural dialogue.

- *Balancing Parameters for Task Optimization*: A well-calibrated LLM typically uses a combination of these parameters to optimize for specific tasks. For example, conversational agents that need to be informative

but engaging may use a moderate temperature (around 0.7), a higher p-value (e.g., 0.95) to encourage nuanced responses, and a k-value of around 20 to maintain contextual relevance. For technical or fact-based responses, lower temperatures (close to 0.5), smaller p-values, and lower k-values may be ideal, producing focused, reliable answers.

By tuning these decoding parameters, we can significantly improve alignment with user expectations, ensuring the model produces responses that are both reliable and contextually appropriate, enhancing its overall effectiveness and user satisfaction.

### 3.3 Model Architecture

The architecture of modern LLMs closely resembles that of the original decoder-only Transformer model introduced by Vaswani et al. (2017), with some minor adjustments. The fact that this foundational architecture has remained largely unchanged highlights its strong performance and adaptability for a wide range of language tasks. However, there is growing evidence that certain architectural choices within the Transformer framework may contribute to issues like hallucination and are not fully optimized for LLMs' current use cases.

**Self-Attention** One key factor contributing to hallucinations is the ***self-attention mechanism***, which is both a strength and a limitation of the Transformer architecture. Self-attention enables the model to capture contextual dependencies by weighting the relevance of all tokens in a sequence to each other, but it lacks a structured mechanism to ensure consistency across longer text spans. As a result, models may generate plausible-sounding but inaccurate information, especially in cases where maintaining factual consistency over multiple sentences or paragraphs is critical. This limitation arises partly because the Transformer treats each token independently, and without explicit grounding, it cannot verify information accuracy. Recent studies have identified a correlation between unpredictable errors in reasoning and the self-attention mechanism, referring to these errors as *attention glitches* Liu et al., 2023. These glitches occur when the model fails to accurately select, extrapolate, or propagate precise information from one part of the context to the next token prediction. This issue is often linked to the "soft-attention" mechanism, where the use of Softmax activation in attention calculations gradually introduces noise, diluting the focus and fidelity of the information as it is processed over long contexts. Consequently, key details may be lost or distorted, contributing to reasoning errors and making the model more prone to hallucinations(Chiang and Cholak, 2022).

**Mitigation** According to Liu et al. (2023), applying a regularizer to "sharpen" attention can help mitigate the phenomenon of attention glitches. They propose that "hard-attention"—where output weights are discrete rather than continuous—would be a more effective approach for the Transformer's attention layer. However, issues with differentiability make this option less computationally efficient, posing practical challenges.

While some architectural weaknesses contributing to hallucinations in Transformers have been identified, these flaws are relatively limited and often challenging to pinpoint due to the complex, layered nature of Transformer models. The structure of Transformers, particularly their reliance on deeply stacked, interdependent layers, makes it difficult to isolate specific sources of error without impacting other functional components. This interconnected design means that even minor modifications can produce unpredictable effects throughout the model, complicating efforts to systematically reduce hallucinations. Consequently, consistent architectural improvements aimed at overcoming hallucinations remain a high research priority, as researchers seek methods to refine the core Transformer structure in ways that enhance reliability without compromising the model's foundational strengths.

# 4 Detection

In this section, we are now reviewing existing approaches to hallucination detection, from simpler, less efficient techniques to more advanced and promising methods, to offer a comprehensive overview of the current state-of-the-art landscape. As explained in Section 3, we are far from developing LLMs that do not suffer from the hallucinations issue, for this reason, detecting them is crucial to prevent misinformation and to establish trust in LLM applications. Existing hallucination detection techniques vary widely in complexity and effectiveness, we are now reviewing some techniques identified by Huang et al. (2023).

## 4.1 Basic Detection Approaches

**N-gram Overlap Metrics** The first basic method to detect hallucinations would be to compare existing metrics both on the prompt and the generated text. N-gram overlap metrics, such as ROUGE (Lin, 2004) and PARENT-T (Wang et al., 2020), are foundational techniques that measure the similarity between generated text and a reference. While these metrics are computationally efficient and widely used in natural language generation tasks, they are often inadequate for detecting hallucinations, as they focus solely on surface-level word matching. This limitation makes them less effective for identifying factual inconsistencies, as these methods do not account for semantic correctness (Maynez et al., 2020).

**Entity and Relation Overlap** *Entity-based* (Nan et al., 2021) metrics aim to improve upon n-gram overlap by checking the presence and accuracy of key entities in the generated text, so in specific cases, where accurate inclusion of entities (e.g., names, places) is crucial, this approach could be informative. *Relation overlap* (Goodrich et al., 2019) methods take this further by examining the relationships between entities, providing a more refined assessment of hallucination in generated content. However, these techniques are still limited when the generated output contains complex information that may not directly match the source text.

## 4.2 Intermediate Detection Techniques

**Classifier-Based Detection** Classifier-based techniques use machine learning models trained to distinguish between hallucinated and accurate content (Maynez et al., 2020). By leveraging labeled datasets classifiers can identify factual inconsistencies and other hallucination markers. Despite their effectiveness, these methods are constrained by the need for large, high-quality labeled datasets and may struggle with domain-specific hallucinations where labeled data is scarce.

**Question-Answering (QA) Based Metrics** QA-based methods involve generating questions from the model's output and attempting to answer them using the source content as a reference (Scialom et al., 2021; Honovich et al., 2021). By comparing the answers generated from the source and the model's output, this method can assess the alignment of information. QA-based approaches offer a more robust assessment than n-gram and entity-based methods, as they focus on the factuality of key information. However, they are computationally intensive and require sophisticated question-generation models, which may introduce additional complexity.

## 4.3 Advanced Detection Techniques

**Uncertainty Estimation and Self-Consistency Checking** Uncertainty estimation techniques use internal model metrics, such as token probabilities, entropy, or response consistency, to gauge the likelihood of hallucination. Farquhar et al. (2024) employed entropy-based methods to detect a subset of hallucinations— confabulations — by letting the model generate multiple answers to a single question, and compare those answer to estimate the entropy measure. Uncertainty estimation offers advantages in detecting subtle hallucinations that surface in response variation. Although promising, these methods can be computationally costly, especially for larger LLMs, as they require generating and comparing multiple outputs for a single prompt.

**Multi-Agent Cross-Examination** A newer approach, multi-agent cross-examination, leverages interactions between two or more LLMs to identify hallucinations (Cohen et al., 2023). This technique, inspired by legal and investigative interviews, employs one model (the "examiner") to query another model (the "examinee") and reveal inconsistencies through multi-turn interactions. By involving multiple LLMs, this approach provides a robust check on factuality, with promising applications in high-stakes environments where accuracy is paramount. However, the added complexity of managing multiple agents and interpreting their interactions poses practical challenges.

### 4.4 External KB Detection Methods

**Retrieval-Based Fact-Checking** Retrieval-based fact-checking combines model output with external, trusted sources to verify the accuracy of generated content (Lewis et al., 2020). This approach retrieves relevant documents or database entries to cross-reference with LLM outputs. Using real-time or up-to-date sources, such as online encyclopedias or domain-specific knowledge bases, it can achieve high accuracy. Retrieval-based methods are particularly useful for long-form or complex hallucinations, which simpler techniques may not catch. These methods, however, require careful implementation of retrieval algorithms and are limited by the quality and availability of external data sources.

## 5 Future directions

In Section 3, we reviewed a variety of studies focused on identifying factors associated with hallucination in language models. A particularly promising area of research is the *explainability of transformers*. Recent studies are increasingly focused on examining the internal states of transformers, aiming to interpret these states to uncover the "reasoning" process that language models undergo when generating each token. This approach seeks to make the model's token generation process more transparent and understandable. For example, studies such as Lee et al. (2024); Hanna et al. (2023) have explored reasoning processes within a controlled, fundamental setting: basic arithmetic tasks. This setting, particularly tasks like addition, proves valuable for examining reasoning because it mirrors key elements of general reasoning processes. In an addition task, for instance, the model must calculate a result based on preceding tokens, retain and transfer information across multiple generation steps (such as carrying over values), and then compute results that build upon this transferred information. This structured, step-by-step reasoning makes basic arithmetic a useful proxy for studying how language models manage sequential, logical dependencies in token generation since it avoids the nuances of natural language meaning, offer simpler, deterministic results that are straightforward to verify, and involve a finite, small set of tokens, making the model's intermediate steps easier to interpret and analyze.

## 6 Conclusion

In this work, various definitions of hallucinations were explored, along with their causes and the most recent techniques used to detect and mitigate them. The journey into this topic has revealed a landscape where challenges persist, but progress is clearly visible. Even though the issue of hallucinations remains unsolved, each new approach brings the field closer to building more reliable models that can better balance accuracy with creative output.

One of the key takeaway of this study is that the problem of hallucinations is multi-faceted. The phenomenon can stem from a range of factors — data quality, model architecture, training processes, and even the inherent ambiguity of human language itself. By identifying these factors, researchers have been able to design detection methods that not only pinpoint when a model might be producing misleading or fabricated outputs but also offer ways to minimize such occurrences.

Newer models continue to being released and many of them are built using open-source frameworks and refined data. This trend has played a significant role in elevating the standard for trustworthy language models, since open-source LLMs let researcher from all over the world investigate the models in their entirety, leading to new and exciting findings on how these complex models are working.

Another important aspect is the shift toward using more advanced training methods. Recent models have been designed to integrate updated techniques that adapt to new challenges as they arise. This means that even if a complete solution to hallucinations is still out of reach, continuous improvements are being made. Each innovation — whether it's a smarter algorithm for detecting anomalies in output or a refined training protocol — adds another layer of protection against misleading results. Such advancements are setting a new benchmark in the realm of language models, where trustworthiness is as prized as creative capability.

It is also worth noting that the ongoing research into hallucinations reflects a broader trend in the field of artificial intelligence. The focus is shifting from simply making models more powerful to making them more interpretable and transparent. This evolution is important because it ensures that the potential for creative expression does not come at the cost of factual reliability. In many ways, the current efforts serve as a bridge between raw performance and practical usability, opening up new possibilities for applications in various fields while also keeping a cautious eye on the limits of current technology.

## References


American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. text revision edition, 2022. doi: 10.1176/appi.books.9780890425787. URL https://www.psychiatry.org/psychiatrists/practice/dsm.

Amos Azaria and Tom M. Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.68. URL https://doi.org/10.18653/v1/2023.findings-emnlp.68.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7654–7664. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.527. URL https://doi.org/10.18653/v1/2022.acl-long.527.

Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong C. Park. Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3145–3157. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.207. URL https://doi.org/10.18653/v1/2023.findings-emnlp.207.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: detecting factual errors via cross examination. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12621–12640. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.778. URL https://doi.org/10.18653/v1/2023.emnlp-main.778.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and

Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2197–2214. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.168. URL https://doi.org/10.18653/v1/2021.emnlp-main.168.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. doi: 10.1038/S41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 864–870. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.76. URL https://doi.org/10.18653/v1/2020.findings-emnlp.76.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM, 2019. doi: 10.1145/3292500.3330955. URL https://doi.org/10.1145/3292500.3330955.

Dirk Groeneveld, Iz Beltagy, Evan Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.841. URL https://doi.org/10.18653/v1/2024.acl-long.841.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *CoRR*, abs/2306.11644, 2023. doi: 10.48550/ARXIV.2306.11644. URL https://doi.org/10.48550/arXiv.2306.11644.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data. *CoRR*, abs/2205.10487, 2022. doi: 10.48550/ARXIV.2205.10487. URL https://doi.org/10.48550/arXiv.2205.10487.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. $Q^2$: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *CoRR*, abs/2104.08202, 2021. URL https://arxiv.org/abs/2104.08202.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023. doi: 10.48550/ARXIV.2311.05232. URL https://doi.org/10.48550/arXiv.2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12): 248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *CoRR*, abs/2307.10169, 2023. doi: 10.48550/ARXIV.2307.10169. URL https://doi.org/10.48550/arXiv.2307.10169.

Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias in llms, 2023. URL https://arxiv.org/abs/2308.14921.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.577. URL https://doi.org/10.18653/v1/2022.acl-long.577.

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=dsUB4bst9S.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention glitches with flip-flop language modeling. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/510ad3018bbdc5b6e3b10646e2e35771-Abstract-Conference.html.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.173. URL https://doi.org/10.18653/v1/2020.acl-main.173.

Mathias Müller, Annette Rios, and Rico Sennrich. Domain robustness in neural machine translation. In Michael J. Denkowski and Christian Federmann, editors, *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020, Virtual, October 6-9, 2020*, pages 151–164. Association for Machine Translation in the Americas, 2020. URL https://aclanthology.org/2020.amta-research.14/.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathy McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2727–2733. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.235. URL https://doi.org/10.18653/v1/2021.eacl-main.235.

OpenAI, Nov 2022. URL https://openai.com/index/chatgpt/.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick,

Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1172–1183. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.92. URL https://doi.org/10.18653/v1/2021.naacl-main.92.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL https://doi.org/10.48550/arXiv.2408.00118.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. Questeval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.529. URL https://doi.org/10.18653/v1/2021.emnlp-main.529.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar,

Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15725–15788. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.840. URL https://doi.org/10.18653/v1/2024.acl-long.840.

Sam Toyer, Olivia Watkins, Ethan Adrian Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, Alan Ritter, and Stuart Russell. Tensor trust: Interpretable prompt injection attacks from an online game. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=fsW7wJGLBd.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521, 2023. doi: 10.48550/ARXIV.2310.07521. URL https://doi.org/10.48550/arXiv.2310.07521.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1072–1086. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.101. URL https://doi.org/10.18653/v1/2020.acl-main.101.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.494/.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. *CoRR*, abs/2004.10450, 2020. URL https://arxiv.org/abs/2004.10450.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.