Hallucination Detection in Machine Generated Text: A Survey

Ashita Saxena, Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Bombay, India
{ashitasaxena, pb}@cse.iitb.ac.in

Abstract

In recent years, the development and deployment of large language models (LLMs) have revolutionized the field of natural language processing. However, these models are prone to generating hallucinations, i.e., outputs that are factually incorrect or inconsistent with the given context. This survey paper provides a comprehensive overview of hallucination detection in machine-generated text. We classify hallucinations into intrinsic and extrinsic types and further distinguish between factuality and faithfulness hallucinations. We explore the origins of hallucinations, identifying key phases such as training data, training phase, and inference. Various approaches for detecting hallucinations are reviewed, including fact overlap-based, entailment-based, weakly supervised classifier-based, question-answering based, retrieval-based, uncertainty-based, prompting-based, and gaze-based methods. Additionally, we examine benchmarks used for evaluating hallucination detection methods, such as FACTOR, FreshQA, MedHALT, HaluEval, and FELM. The paper also delves into human cognitive behavior and its relevance to hallucination detection, highlighting basic terminologies and experimental designs in eye tracking. In conclusion, we discuss future directions for research, emphasizing the need for improved self-correction mechanisms, understanding of knowledge boundaries, and balancing creativity with truthfulness in LLM outputs. This survey aims to provide a foundation for further research and development in the field of hallucination detection in LLMs.

1 Introduction

Hallucination detection in text refers to the task of identifying and validating information that is inaccurately or falsely represented within textual content. Detection of hallucination involves examining the claims made in the text and assessing their alignment with the surrounding context and external knowledge. Addressing hallucinations has become paramount, particularly in the context of automatically generated text utilizing powerful language models (LLMs), which often exhibit humanlike fluency but are prone to hallucinatory outputs (Zhang et al., 2023a; Alkaissi and McFarlane, 2023). Natural Language Generation (NLG) has made tremendous progress in neural text generation with the advent of large pre-trained language models like BERT (Devlin et al., 2018) and the GPT series (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023). Although text generation using these models is fluent, it is often observed that the generated text is divergent or unfaithful to the source text (Kryściński et al., 2019a; Wiseman et al., 2017; Dhingra et al., 2019). This problem of generating contradicting or irrelevant text is termed *hallucination* (Maynez et al., 2020b).

2 Motivation

Recent advancements in artificial intelligence, especially in large language models (LLMs) (Agrawal et al., 2022; Radford et al., 2019), have revolutionized various fields, including healthcare and legal domains (Singhal et al., 2023a; Bernsohn et al., 2024; Wang et al., 2023b). These models, capable of comprehending and generating human-like text by learning from extensive text data, serve as valuable tools for medical professionals, legal experts, researchers, and students (Singhal et al., 2023b; Nguyen et al., 2024; Louis et al., 2024; Wiratunga et al., 2024). However, despite their impressive capabilities, LLMs face unique challenges such as hallucination (Ji et al., 2023; Bang et al., 2023), where they produce plausible yet incorrect or unverified information.

To address these challenges, it is essential to develop methods to evaluate and mitigate hallucinations. Central to this effort is the creation of robust datasets specifically designed for hallucination detection. These datasets can facilitate the identi-

fication and reduction of hallucinations in LLM outputs, enhancing the reliability of these models in high-stakes fields. Furthermore, improving the transparency and interpretability of LLMs can help users understand the limitations and potential inaccuracies of the generated content. By improving our ability to detect and mitigate hallucinations, we can ensure that LLMs are safer and more reliable tools in both medical and legal contexts, ultimately protecting the welfare of individuals and the integrity of critical decision-making processes.

Moreover, the development of hallucination detection techniques can drive advancements in the broader field of NLP. It encourages the research community to focus on building models that are not only fluent and coherent but also factually accurate and trustworthy. This shift in focus can lead to the creation of more robust and reliable AI systems, capable of being deployed in a wider range of applications with confidence. Addressing hallucinations aligns with the ethical considerations of AI development, ensuring that these powerful technologies are used responsibly and beneficially for society.

3 Background and Definitions

The term hallucination was inspired by psychology. In the medical context, hallucinations refer to the particular type of perception realized by an individual without any external stimulus (Blom, 2010). Hallucination, as a psychological term, refers to an unreal perception that looks real on the surface. In the same way, in NLG, the generated text may contain information that might look correct but if we verify the information present, it might contain unfaithful or illogical text.

Various works use different categorizations of hallucination in NLP tasks. In this survey, we will see two types of categorizations.

Maynez et al. (2020a) first introduced the division of hallucination into **intrinsic** and **extrinsic hallucination**. Recently, a more fine-grained categorization of hallucination was introduced by Huang et al. (2023). They divide hallucinations into two broad categories: **Factuality Hallucinations** and **Faithfulness Hallucinations**. Factuality hallucinations are further divided into two subcategories: Factual Inconsistency and Factual Fabrication. Faithfulness hallucinations are further divided into three sub-categories: Instruction Inconsistency, Context Inconsistency and Logical Inconsistency. These two types of classifications are discussed in detail in the following sections.

3.1 Types of Hallucination: Intrinsic and Extrinsic

3.1.1 Intrinsic Hallucination

Intrinsic Hallucinations occur when the output generated by any NLG model contradicts the source text. For example, in a machine translation task, intrinsic hallucinations are defined as a span of the word(s) in the generated sequence containing incorrect information but they might represent the same entity type. Similarly, in the summarization task, if the generated summary contradicts the given source information or document, it is referred to as intrinsic hallucination.

3.1.2 Extrinsic Hallucination

Extrinsic Hallucinations occur when the output generated by any NLG model cannot be verified by the source information. In other words, the generated output neither contradicts nor is supported by the source information. It is important to note that extrinsic hallucinations are challenging to detect as they are not implied by the source text or information. one interesting identity of extrinsic hallucinations is that it does not always contain factually incorrect data i.e. although the generated output text might not be validated whether it is true or false from the source information provided. But, the generated output can be factually correct considering the external or world knowledge. But, identifying extrinsic hallucination improves the consistency with the reference text and further identifies the content which is not required for the given specific task or context. The example of extrinsic hallucination in machine translation refers to the span of words consisting of additional information which can't be inferred from the given input or source text. In the context of the summarization task, extrinsic hallucinations refer to the output text neither supported nor contradicts by the given input article

The definitions of intrinsic and extrinsic slightly vary depending on the task, for example in machine translation intrinsic hallucination refers to the substitution of some other entity in place of the real or true entity while in abstractive summarization intrinsic hallucination refers to the contradiction to the source text. For other NLP tasks, a few examples are shown in Figure 1 taken from Ji et al. (2022).

Task	Sub-Task	Type	Source	Output
Abstractive Summarization [133]		Intrinsic	The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.	The first Ebola vaccine was approved in **2021**.
		Extrinsic		**China** has already started clinical trials of the COVID-19 vaccine.
Dialogue	Task-oriented [103]	Intrinsic	inform (NAME = pickwick hotel, PRICERANGE = moderate)	the hotel named pickwick hotel is in a **high** price range.
		Extrinsic		the pickwick hotel **in san diego** is a moderate price range
	Open-domain [154]	Intrinsic	Dialog History: Speaker 1: What do you think about Murray? Speaker 2: I think Murray is a great player he just needs to stay healthy in order to compete more. Who do you like best?	Speaker 1: I like **Roger Nadal** the most!
		Extrinsic	Knowledge: In tennis, the term Big Four refers to the quartet of men's singles players comprising Roger Federer, Rafael Nadal, Novak Djokovic, and Andy Murray.	Speaker 1: I like Djokovic. He has played **in the top ten singles players of the world**.
Generative QA [129]		Intrinsic	Question: dow jones industrial average please?	Answer: The Dow Jones Industrial Average(DJIA) is an index of **30 major U.S.stock indexes**. The DJIAwas created in 1896 by Charles Dow, a statistician and editor of The Wall Street Journal. The index was first published in the mid-1880s, and **was last updated on May 26,1896.**
		Extrinsic	Question: definition of sadduction	Answer: Sadducees. noun. **The definition of a sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power.**1 a member of a Jewish sect that was active during the Second Temple.
Data2text [195]		Intrinsic	TEAM CITY WIN LOSS PTS FG_PCT BLK Rockets Houston 18 5 108 44 7	The Houston Rockets **(18-4)** defeated the Denver Nuggets (10-13) 108-96 on Saturday.
		Extrinsic	Nuggets Denver 10 13 96 38 7	**Houston has won two straight games and six of their last seven.**
Translation [224]		Intrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	**Jerry didn't go** to the bookstore.
		Extrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	Michael **happily** went to the bookstore **with his friend.**

Figure 1: Examples of Intrinsic and Extrinsic hallucinations for different NLG tasks

3.2 Types of Hallucination: Factuality and Faithfulness

The concept of hallucination originates from the fields of pathology and psychology, where it *denotes the perception of an entity or event that does not exist in reality* (Macpherson and Platchias, 2013). In natural language processing (NLP), hallucination describes a phenomenon *where generated content appears nonsensical or unfaithful to the source material* (Filippova, 2020; Maynez et al., 2020a). This concept loosely mirrors the hallucinations observed in human psychology. Hallucinations in natural language generation can generally be divided into two main types: intrinsic and extrinsic (Cao et al., 2021; Li et al., 2022; Ji et al., 2023). Intrinsic hallucinations involve outputs that contradict or conflict with the source content, while extrinsic hallucinations involve outputs that cannot be verified or supported by the source content.

In the era of large language models (LLMs), their versatile capabilities have led to widespread application across various fields, revealing limitations in traditional task-specific categorization paradigms. Since LLMs focus on user-centric interactions and alignment with user directives, and their hallucinations primarily occur at factual levels, a more detailed taxonomy is proposed by Huang et al. (2023), building on the foundational work of Ji et al. (2023). Their refined taxonomy aims to capture the distinct complexities associated with LLM hallucinations. To illustrate their definition of LLM hallucination more intuitively, examples for each type of hallucination are provided in Table 1 and Table 2, along with corresponding explanations. Figure 2 shows some intuitive examples of factuality and faithfulness hallucinations (taken from Huang et al. (2023)). The details of their proposed categories are elaborated below:

3.2.1 Factuality Hallucination

Current LLMs sometimes generate outputs that are either factually incorrect or potentially misleading, which undermines the reliability of artificial intelligence. These factual inaccuracies are referred to as factuality hallucinations. Based on whether the generated factual content can be corroborated with a reliable source, these hallucinations can be classified into two main types:

- **Factual Inconsistency** occurs when the out-

put of an LLM contains facts that, despite being grounded in real-world information, present contradictions. This type of hallucination is the most common and stems from various sources, including the LLM's acquisition, storage, and expression of factual knowledge. For instance, as illustrated in Table 1, when asked about "the first person to land on the Moon," the model incorrectly generated "Yuri Gagarin" which contradicts the actual historical fact.

- **Factual Fabrication** refers to cases where the LLM produces facts that cannot be verified against established real-world knowledge. For example, as shown in Table 1, while "the origins of unicorns" lack empirical evidence, the model invented a plausible historical origin for unicorns.

3.2.2 Faithfulness Hallucination

As LLMs' applications become more user-centric, maintaining consistency with the instructions and contextual information provided by users is crucial. Additionally, the faithfulness of an LLM is evident in the logical coherence of its generated content. From this standpoint, three subtypes of faithfulness hallucinations are proposed by Huang et al. (2023):

- **Instructional inconsistency** pertains to instances where the outputs of an LLM diverge from a user's directive. As illustrated in Table 2, the user's intent is translation; however, the LLM erroneously veered from this instruction and instead performed a question-answering task.

- **Context inconsistency** arises when the output of an LLM fails to adhere to the contextual information provided by the user. For instance, as depicted in Table 2, although the user specified that the source of the Nile is in the Great Lakes region of central Africa, the LLM's response contradicted this context.

- **Logical inconsistency** occurs when the outputs of an LLM exhibit internal logical contradictions. For example, as demonstrated in Table 2, while the reasoning step of dividing both sides of the equation by 2 is accurate, the final answer of x=4 contradicts the reasoning chain, resulting in an incorrect outcome.

4 Origin of Hallucination

Language models can generate hallucinated or fabricated content due to various factors stemming from how they acquire knowledge and capabilities. This section examines the underlying reasons behind hallucinations in large language models, broadly grouped into three main areas: the data used for training, the training process itself, and the inference or generation stage.

4.1 Hallucinations Originating from Training Data

The pre-training data forms the foundation for large language models, allowing them to develop general capabilities and acquire factual knowledge (Zhou et al., 2024). However, this training data can inadvertently become a source of hallucinated or fabricated content generated by the model. This issue primarily stems from two factors: potential pitfalls arising from flawed data sources and suboptimal utilization of the factual knowledge present in the data.

Flawed Data Sources While increasing the scale of pre-training data significantly enhances language models' competencies (Kaplan et al., 2020; Hoffmann et al., 2022) , maintaining consistent data quality remains a challenge, potentially introducing misinformation and biases (Bender et al., 2021; Weidinger et al., 2021). Furthermore, the absence of specific domain knowledge and up-to-date facts in the training data can create knowledge boundaries for the language model, limiting its performance in certain scenarios.

Inferior Data Utilization The pre-training data contains a rich repository of real-world factual information, allowing large language models to absorb and encode a vast amount of factual knowledge within their parameters (Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020). However, despite possessing this extensive knowledge base, language models can still generate hallucinations containing factual inaccuracies due to suboptimal utilization and integration of the parametric knowledge they have acquired.

4.2 Hallucination Originating in Training Phase

The pre-training stage lays the foundation for large language models, typically employing a transformer-based architecture to perform causal language modeling on vast text corpora. How-

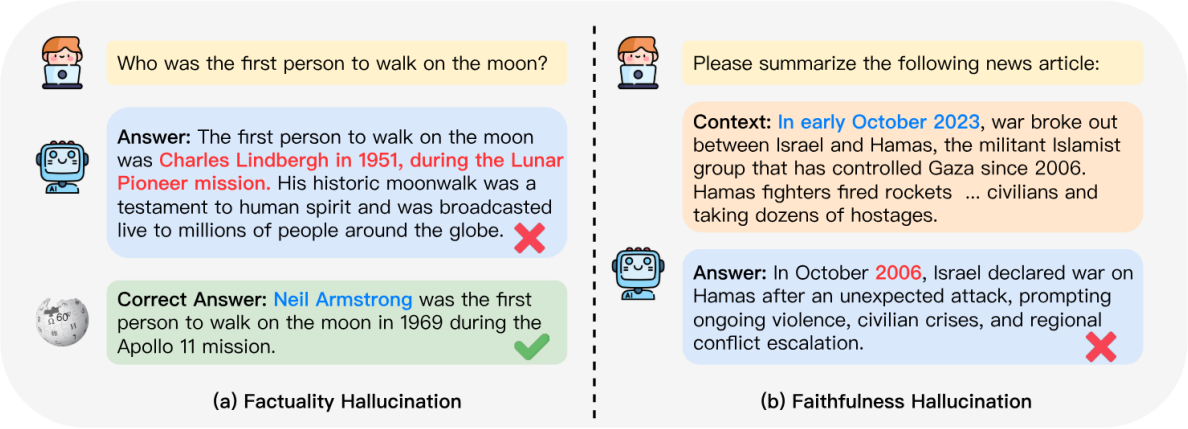

Figure 2: Intuitive examples of factuality and faithfulness hallucinations.

Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Factuality Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon.	The LLM's response is factual inconsistency as Neil Armstrong was the first person to land on the Moon, not Yuri Gagarin, who was the first human in space.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since there's no verified, real-world evidence supporting the existence of unicorns in Atlantis, especially in connection with royalty.

Table 1: Examples of Faithfulness Hallucinations. Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination. These examples are taken from Huang et al. (2023)

ever, certain aspects of the architectural design and training strategies employed during this stage can potentially lead to hallucinated outputs. Most large language models adopt a transformer-based architecture following the paradigm established by GPT (Radford et al., 2018; Brown et al., 2020), where they learn representations through a causal language modeling objective. This approach is exemplified by models such as OPT (Zhang et al., 2022), Falcon (Penedo et al., 2023), and Llama-2 (Touvron et al., 2023). Despite its success, this unidirectional modeling approach has certain drawbacks, primarily related to Inadequate Unidirectional Representation and Attention Mechanisms. The unidirectional nature of this modeling, while facilitating efficient training, also presents limitations. By exclusively utilizing context from a single direction, it may hinder the model's ability to capture intricate contextual dependencies, potentially increasing the risk of generating hallucinated content (Li et al., 2023b). Recent research (Liu et al., 2024) has demonstrated that these models can occasionally exhibit unpredictable reasoning errors in the context of algorithmic reasoning tasks, spanning both long-range and short-range dependencies, regardless of model scale. A potential contributing factor is the limitations of soft attention mechanisms (Hahn, 2020; Chiang and Cholak, 2022), where attention may become diluted across positions as sequence length increases.

4.3 Hallucination Originating from Inference

The decoding process is crucial to get the full potential of large language models after they have undergone pre-training. However, certain limitations or flaws within the employed decoding strategies

Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Faithfulness Hallucination	Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. Instead of performing the translation as instructed, the LLM chose to answer the question.
	Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.
	Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, the next step inaccurately divides 8 by 2 to yield a result of x = 3, which is inconsistent with the earlier reasoning.

Table 2: Examples of Faithfulness Hallucinations. Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination. These examples are taken from Huang et al. (2023).

can result in the models generating hallucinated or fabricated content that deviates from factual information.

Currently, large language models predominantly use stochastic sampling (Fan et al., 2018; Holtzman et al., 2019) as their decoding strategy. The motivation behind introducing randomness into these decoding strategies arises from the observation that sequences with high likelihood often result in surprisingly low-quality text, a phenomenon known as the likelihood trap (Stahlberg et al., 2022; Holtzman et al., 2019; Meister et al., 2020; Zhang et al., 2023b). However, the diversity gained from this randomness in decoding comes at a cost, as it is positively correlated with an increased risk of hallucinations (fabricated or nonsensical content) in the generated text (Dziri et al., 2021a; Chuang et al., 2023).

5 Approaches for Hallucination Detection

The ability to identify hallucinations or fabricated content generated by large language models is essential for ensuring the reliability and trustworthiness of their outputs. Following sections discuss various approaches to hallucination detection

5.1 Fact Overlap-based Approaches

Ensuring LLMs are faithful in providing context or user directives is crucial for practical use in various applications such as summarization and interactive dialogue systems. Detection of faithfulness hallucination primarily focuses on aligning the generated content with the given context to avoid extraneous or contradictory output. When evaluating faithfulness, a common method involves measuring the overlap of key facts between the generated and source content. Metrics can be categorized based on entities, relation triples, and knowledge (Lin, 2004; Wang et al., 2020b; Maynez et al., 2020a).

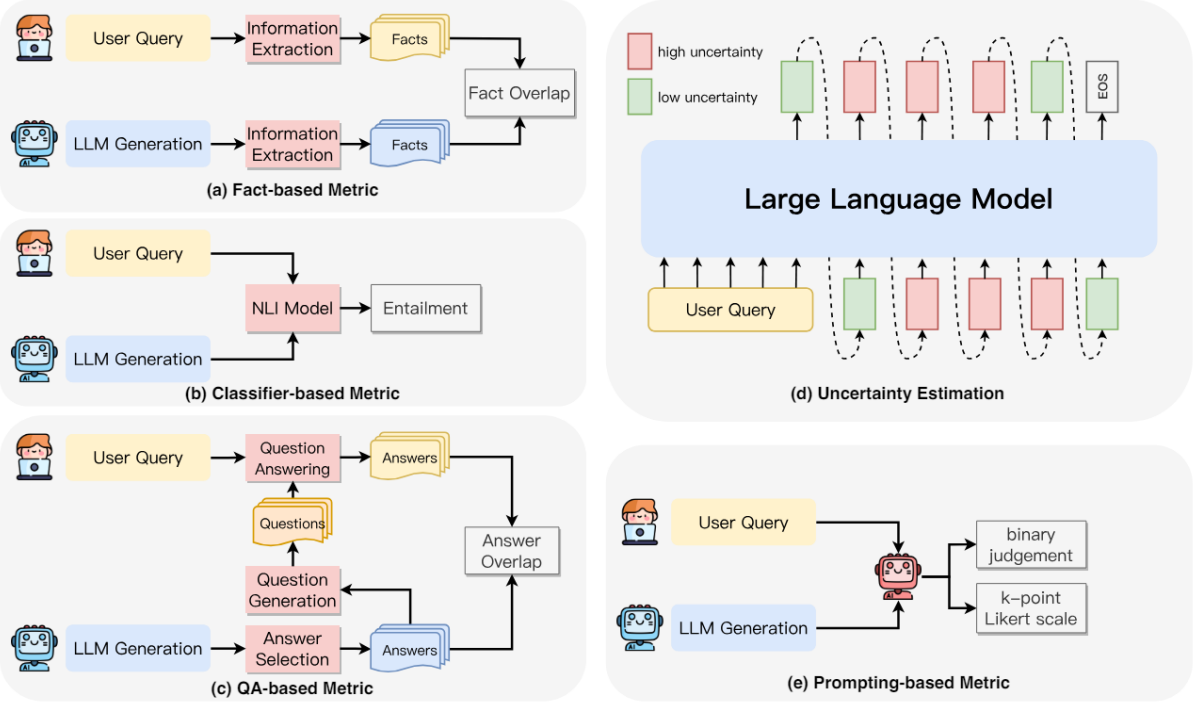

Figure 3: The illustration of detection methods for faithfulness hallucinations: a) Fact-based Metrics, which assesses faithfulness by measuring the overlap of facts between the generated content and the source content; b) Classifier-based Metrics, utilizing trained classifiers to distinguish the level of entailment between the generated content and the source content; c) QA-based Metrics, employing question-answering systems to validate the consistency of information between the source content and the generated content; d) Uncertainty Estimation, which assesses faithfulness by measuring the model's confidence in its generated outputs; e) Prompting-based Metrics, wherein LLMs are induced to serve as evaluators, assessing the faithfulness of generated content through specific prompting strategies.

N-gram based: Evaluation metrics like ROUGE (Lin, 2004) and PARENT-T (Wang et al., 2020b) can assess faithfulness by treating the source content as the reference. However, these metrics show poor correlation with humans due to language diversity and reliance on surface-level matching (Maynez et al., 2020a).

Entity-based: Metrics that focus on entity overlap are commonly used in summarization tasks to ensure accurate summaries. A metric introduced by (Nan et al., 2021) quantifies entity hallucination by measuring the precision of named-entities in the summary against the source entities.

Relation-based: (Goodrich et al., 2019) focus on the overlap of relation tuples and introduce a metric that calculates the overlap of relation tuples extracted using trained end-to-end fact extraction models.

Knowledge-based: The Knowledge F1 metric introduced by (Shuster et al., 2021) assesses how well the model's generation aligns with the provided knowledge in knowledge-grounded dialogue tasks.

5.2 Entailment-based Approaches

Using Natural Language Inference (NLI) to assess text trustworthiness is a prevalent concept. Studies (Falke et al., 2019; Maynez et al., 2020a) have used NLI datasets to spot inaccuracies, particularly in abstract summarization. However, Mishra et al. (2021) noted limitations in detecting inconsistencies due to input granularity discrepancies. Advanced studies suggest methods like fine-tuning on adversarial datasets (Barrantes et al., 2020), breaking down entailment decisions at the dependency arc level (Goyal and Durrett, 2020), and segmenting documents into sentence units for improved accuracy in hallucination detection.

5.3 Weakly Supervised Classifier-based Approaches

Leveraging data from related tasks to fine-tune classifiers has potential but acknowledges gaps between tasks. Addressing this, Kryściński et al. (2019b) introduced rule-based transformations to

create weakly supervised data for refining classifiers. Zhou et al. (2020) developed a method for token-level hallucination detection. Dziri et al. (2021b) created adversarial synthetic data, while Santhanam et al. (2021) focused on factual consistency in dialogue tasks.

5.4 Question-Answering Based Approaches

Metrics based on Question-Answering are gaining attention for capturing information overlap between a model's output and source. These metrics select target answers and generate questions to assess reliability. Implementations include (Durmus et al., 2020; Wang et al., 2020a; Scialom et al., 2021; Honovich et al., 2021), showing varied performance outcomes.

5.5 Retrieval-based Approaches

One strategy to detect factual inaccuracies in LLM outputs is to compare the generated content against reliable knowledge sources, aligning with fact-checking workflows. However, traditional fact-checking methods often make simplifying assumptions, leading to limitations in complex real-world scenarios. Recent approaches address this by incorporating components like claim decomposition, uncurated web evidence retrieval, claim-focused summarization, and veracity classification (Chen et al., 2023). Techniques also aim to resolve conflicting evidence (Galitsky, 2023) and compute fine-grained factual scores for long-form generations by decomposing them into atomic facts and checking against knowledge sources (Min et al., 2023).

5.6 Uncertainty-based Approaches

Several approaches aim to detect hallucinations without relying on external knowledge sources, operating in zero-resource settings. These methods are based on the premise that hallucinations stem from the model's uncertainty. By estimating the uncertainty of the factual content generated, it becomes possible to identify hallucinations without the need for evidence retrieval. The internal states of large language models can indicate their uncertainty, manifested through metrics like token probability or entropy. Varshney et al. (2023) determine model uncertainty for key concepts by considering the minimum token probability within those concepts, where lower probabilities signal higher uncertainty. Luo et al. (2023a) propose a self-evaluation approach, grounding on the idea that a model's ability to reconstruct a concept from its own explanation reflects its proficiency with that concept, thus indicating uncertainty.

When systems can only be accessed through API calls, access to the output's detailed probability distribution may be limited. To address this, recent studies have focused on exploring a model's uncertainty using methods such as natural language prompts or analyzing its behavior. For example, Manakul et al. (2023) identified hallucinations in a language model by evaluating consistency among responses to the same prompt. Agrawal et al. (2023) suggest using indirect queries, which ask open-ended questions to gather specific information, unlike direct queries that explicitly seek verification. Another approach involves assessing uncertainty by comparing multiple generations of language models. Cohen et al. (2023) proposed the LMvLM method, where one language model questions another to uncover inconsistencies during interactive sessions, inspired by legal cross-examination practices.

Hallucinations in conditional text generation are linked to high model uncertainty. Uncertainty estimation, explored in Bayesian deep learning (Blundell et al., 2015; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017) determines predictive entropy for overall uncertainty. Some studies (Malinin and Gales, 2020) quantify model uncertainty using log probability.

5.7 Prompting-based Approaches

The exceptional ability of LLMs to follow instructions has recently highlighted their potential for automated assessment (Chiang and Lee, 2023; Liu et al., 2023; Wang et al., 2023a). Leveraging this capability, researchers have explored new approaches to evaluating the accuracy of content generated by models (Luo et al., 2023b; Laban et al., 2023; Adlakha et al., 2023; Gao et al., 2023; Jain et al., 2023). By providing clear evaluation guidelines to LLMs and supplying them with both the model-generated and original content, they can effectively evaluate accuracy. The resulting evaluation can be a binary determination of accuracy (Luo et al., 2023b) or a Likert scale with k-points indicating the level of accuracy (Gao et al., 2023). For prompt selection, evaluation prompts can involve direct prompting, chain-of-thought prompting (Adlakha et al., 2023), in-context learning (Jain et al., 2023), or allowing the model to produce evaluation results along with explanations (Laban et al., 2023).

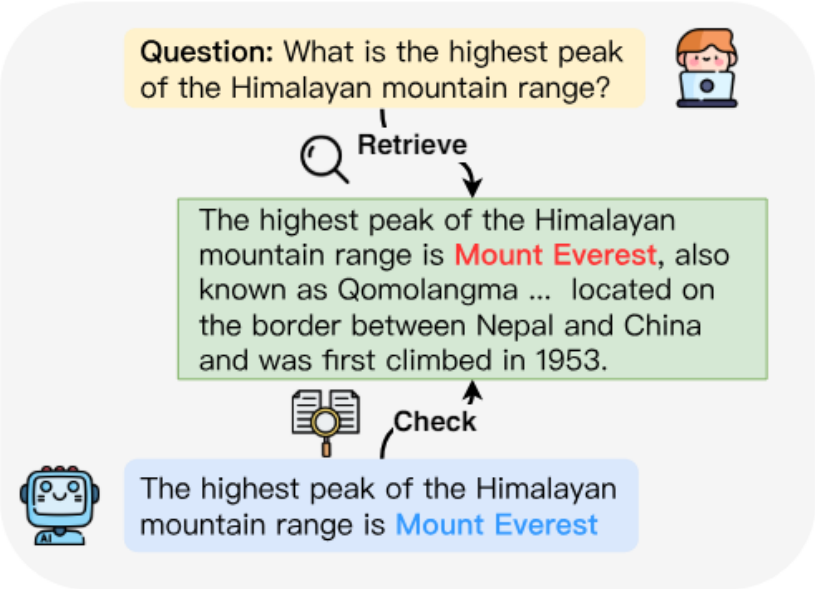

Figure 4: An example of detecting factuality hallucination by retrieving external facts.

5.8 Gaze-based Approaches

In recent years, various attempts have been made to investigate the correlation of human attention with the machine attention of a pre-trained large language model (Eberle et al. (2022), Sood et al. (2020a), Bensemann et al. (2022)). Eberle et al. (2022) highlighted the inability of cognitive models to account for the higher level cognitive activities like semantic role matching, hence motivating the use of large language models (LLMs) for modelling the human gaze. Hollenstein et al. (2021) showed the efficacy of LLMs in predicting the gaze features for multiple languages, including English, Russian, Dutch and German. Barrett et al. (2018) used natural reading eye-tracking corpus for regularizing attention function in a multi-task setting. Sood et al. (2020b) investigates the integration of the gaze-based text saliency model with vanilla transformers (Vaswani et al., 2017) for directly incorporating gaze predictions into the attention mechanisms for paraphrase detection and sentence compression tasks.

6 Hallucination Detection Benchmarks

6.1 FACTOR

Muhlgay et al. (2023) introduced a method to quantitatively evaluate the factuality of language models (LMs) by generating benchmarks through perturbing factual statements from a specified corpus. This approach led to the creation of two benchmarks called Wiki-FACTOR and News-FACTOR. The process involved using prompts with specific error types to guide InstructGPT in generating non-factual completions based on a given prefix text. The resulting responses were then filtered for fluency and self-consistency to form the basis for multi-choice tasks. The evaluation of an LM's factuality was based on whether the model was more likely to produce factually correct completions compared to non-factual ones.

6.2 FreshQA

Vu et al. (2023) introduced FreshQA as a benchmark to assess the factuality of existing large language models (LLMs) by focusing on potential hallucinations stemming from outdated knowledge. This benchmark consisted of 600 manually created questions with answers that could change over time or contain factually incorrect premises. It primarily evaluated LLMs on their ability to handle rapidly changing information and identify questions with false premises. The evaluation process included two modes: RELAXED, which assessed the correctness of the primary answer, and STRICT, which further evaluated the accuracy of all facts within the answer. The factuality of LLMs was judged based on the accuracy of their responses as determined by human annotations.

6.3 Med-HALT

Med-HALT, developed by Umapathi et al. (2023), focuses on the challenges faced by large language models (LLMs) in the medical field, particularly in relation to hallucinations. This benchmark assesses LLMs' reasoning and memory skills in a medical context through multiple-choice questions sourced from various countries. The reasoning task, consisting of 18,866 samples, evaluates LLMs' ability to discern incorrect or irrelevant options and fake questions in medical multiple-choice questions. The memory task, comprising 4,916 samples, assesses LLMs' capacity to recall and accurately generate factual information by linking PubMed abstracts/titles or producing titles from given links and PMIDs. LLM performance is evaluated based on their accuracy in answering test questions or a Pointwise Score that considers both correct answers and penalties for incorrect responses.

6.4 HaluEval

HaluEval, introduced by Li et al. (2023a), aims to evaluate LLMs' ability to recognize hallucinations. This benchmark was created using a combination of automated generation and human annotation, resulting in 5,000 general user queries paired with ChatGPT responses and 30,000 task-specific samples. The automated generation process utilizes a "sampling-then-filtering" approach, drawing from various datasets to sample hallucinated answers and selecting the most plausible ones. Human annotation involves processing Alpaca-sourced queries with ChatGPT to assess the presence of hallucinated content in multiple responses.

6.5 FELM

FELM, developed by Zhao et al. (2024), differs from previous studies by evaluating factuality across five domains: world knowledge, science and technology, mathematics, writing and recommendation, and reasoning. Unlike prior research that induced hallucinations based on specific patterns, this benchmark employs ChatGPT to generate responses in a zero-shot setting, resulting in 817 samples (3948 segments). Each segment is annotated for factuality, error reasons, error type, and external references. Serving as a platform for factuality detectors, the benchmark uses the F1 score and balanced classification accuracy to assess factual errors at both the segment and response levels.

7 Human Cognitive Behaviour and Hallucination

7.1 Basic Terminologies

Reichle et al. (2003) describes various connections of the annotator's gaze behaviour to the reading patterns. We briefly explain three major gaze features and their usage in the context of natural language processing tasks in this section.

7.1.1 Saccades

Contrary to popular belief, reading doesn't really entail the eyes naturally gliding out across text. Instead, saccades—rapid, brief movements—of the eyes are made. Although there are rare exceptions, saccades typically advance the gaze 6 to 9 character spans. Saccades may take 20–50 milliseconds to accomplish, depending on how lengthy the movement is.

In the process of saccadic motion, no information is collected. Saccadic suppression is the term used to describe this phenomenon of decreased susceptibility to visual stimuli (Matin, 1974). This is due to the fact that throughout a saccade, the eyes move so quickly across the stationary visual stimuli that we only see a blur and not new information (Rayner, 1998).

7.1.2 Fixation

Martinez-Conde et al. (2004) defines fixation as the firm focus of gaze on text. It should be observed that even when the sight is fixed, the eyes are constantly moving. Though their magnitude should make them evident to us, we are unaware of such eye movements. If fixational eye movements are blocked for whatever reason—including brain adaptation—our visual perception may completely vanish.

The visual data can only be extracted from the words during fixations. Due to this, normal reading is frequently compared to the viewing of a slide show, when only a few sentences of text are displayed for roughly a second at a time. It's intriguing to note that, like saccade length, the time of the fixation can vary greatly. Fixation typically lasts between 200 and 250 ms (Reichle et al., 2003).

Word length and indeed the amount of space around them appear to have a big impact on where readers decide to focus their attention next in the

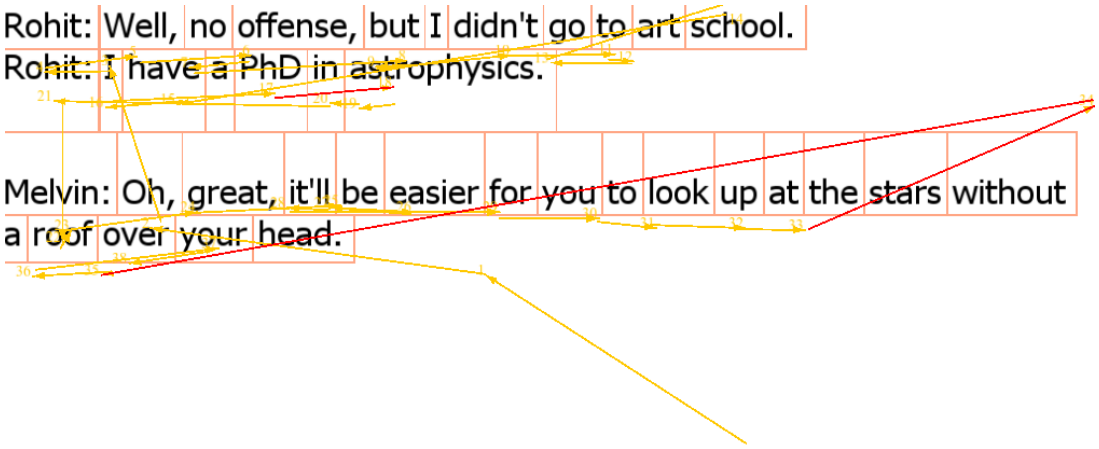

Figure 5: Saccadic Movements

document (Reichle et al., 2003). The preceding hypothesis is supported by a number of further investigations. Rayner (1979) describes the effects of the size of a phrase that also is fixated on the length of saccades. McConkie et al. (1988) investigates the variations in word length-dependent word fixation patterns in readers. Ehrlich and Rayner (1981) explores these patterns. Despite the fact that predicted word is skipped more frequently than unpredictable ones, contextual limitations have minimal effect on the location where a subject's eyes land inside a word.

7.1.3 Pupil Dilation

The phenomenon of pupil enlargement is called pupil dilation. The diameter of the retina's pupil is sensitive to a variety of cognitive functions. Zénon (2019) enlists the possible cognitive scenarios which can directly or indirectly affect pupil diameter which include the following:

1. Mental effort
2. Surprise
3. Emotion
4. Decision Processes
5. Decision Biases
6. Value beliefs
7. Volatility
8. Exploitation Exploration trade-off
9. Attention
10. Uncertainty (Expected and Unexpected)

Based on substantial evidence, (Zénon, 2019) suggests that the updating of internal models in the brain is the fundamental information-theoretic mechanism that underlies the collection of experiences that cause changes in pupil-linked arousal. When a stimulus is presented, the pupillary reaction is proportional to how much information that stimulus contains about it and how much information it offers about other task factors. (Zénon, 2019) tries to define all the above cognitive processes in terms of information gain and reports the similarities between pupillary responses and information gain using KL (Kullback–Leibler) divergence. (Hess and Polt, 1960)) first documented the well-known reversible relationship between emotions and pupil dilation, finding that when individuals looked at painful photos, their pupils shrank, whereas when they glanced at pleasant pictures, their pupils grew.

(Bradley et al., 2008) found that there is a substantial correlation between skin conductance with dilated pupils, suggesting that there could be a separate mechanism behind emotion regulation that primarily involves autonomic modulation of both the dilation muscles. (Bradley et al., 2008) work significantly support the notion that pupillary modifications during picture gazing are transmitted by sympathetic stimulation activity and that pupil dilatates are dictated by emotional response regardless of whether images are pleasant or unpleasant. According to (Hyönä et al., 1995), the difference in the cognitive effort can also be assessed by the vari-

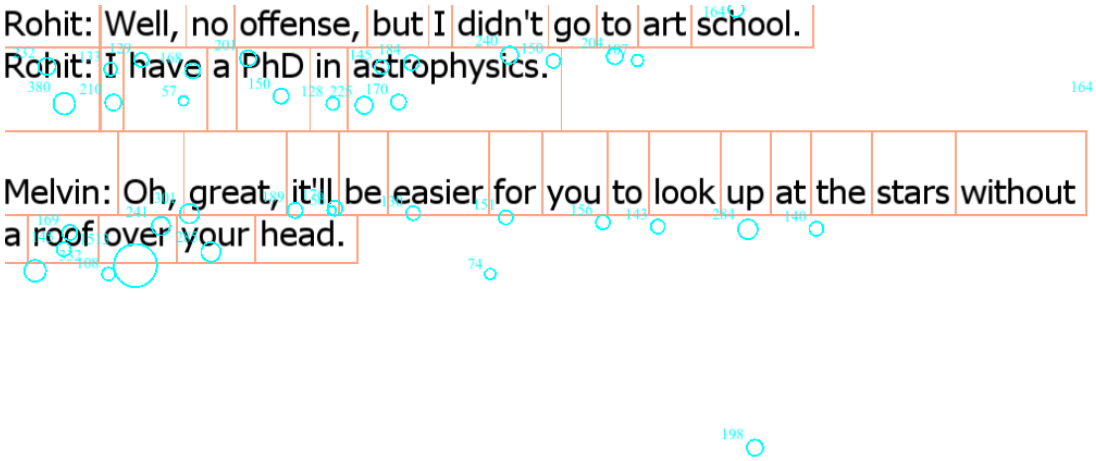

Figure 6: Fixation Points

ation in pupil dilation's magnitude. Besides two significant experiments, the relevance of pupillary response in assessing cognitive function was fully investigated. The first experiment compared the average pupil size's response to simultaneous interpretation to the global cognitive stress of seeing and repeating a text that had been presented orally.

7.2 General Eye Tracking Experiment Design

This section describes a general eye-tracking experiment as described by (Conklin and Pellicer-Sánchez, 2016):

1. **Examine the attributes of the eye-tracking device**: There are several kinds of eye trackers, each with a unique set of parameters that make them more or less suitable for studying various linguistic phenomena. A system must be able to supply the information required to respond to research inquiries. In general, greater sampling rates, monocular recording (rather than binocular), head-supported systems, and/or the use of chin rests result in superior accuracy and resolution. Nevertheless, imprecision is typically not an issue with eye trackers that run at 200Hz. The majority of reading research employs eye trackers with frequencies ranging from 500Hz to 1,000Hz. While devices with lower sample rates can be utilised for reading, the quantity of information required to compensate for that sampling frequency's added imprecision is unfeasible.

2. **Familiarity with the process by which the eye-tracker and related software operate**: It's crucial to be able to calibrate an eye-tracker correctly in order to get reliable data. Data that is not exact will be produced by poor calibration. A nine-point calibration is often performed at the start of an experiment, at extra predetermined times in longer investigations, and so when eye-drifting was present. To ensure that data is being outputted correctly, it is crucial to perform an experiment at least once before the final run.

3. **Choosing the appropriate stimuli**: Critical stimuli must be accurately matched for factors including lexical ambiguity, grammatical structure, word class, length, frequency, predictability, and orthographic uniformity because it has been demonstrated that these factors affect fixation duration. Studies frequently benefit from a control scenario or stimuli that serve as a benchmark. The experimental stimulus and the control stimulus need to be somewhat similar. In order to prevent effects from being driven by diverse contexts, critical stimuli should emerge in situations that are the same or as comparable as feasible.

   Examples of appropriate stimuli are those with the same amount of words, within identical syntactic frames, and equal for bias/predictability. Also, if there is a potential that spillover effects may occur, the region immediately after the crucial stimulus ought to match exactly or be the same. Because reading speed often declines as a reader moves through a book and because words near the end of a sentence and phrases at the conclusion of passages are read more slowly, critical stimulus should be supplied in comparable locations. For instance: New Courier, where

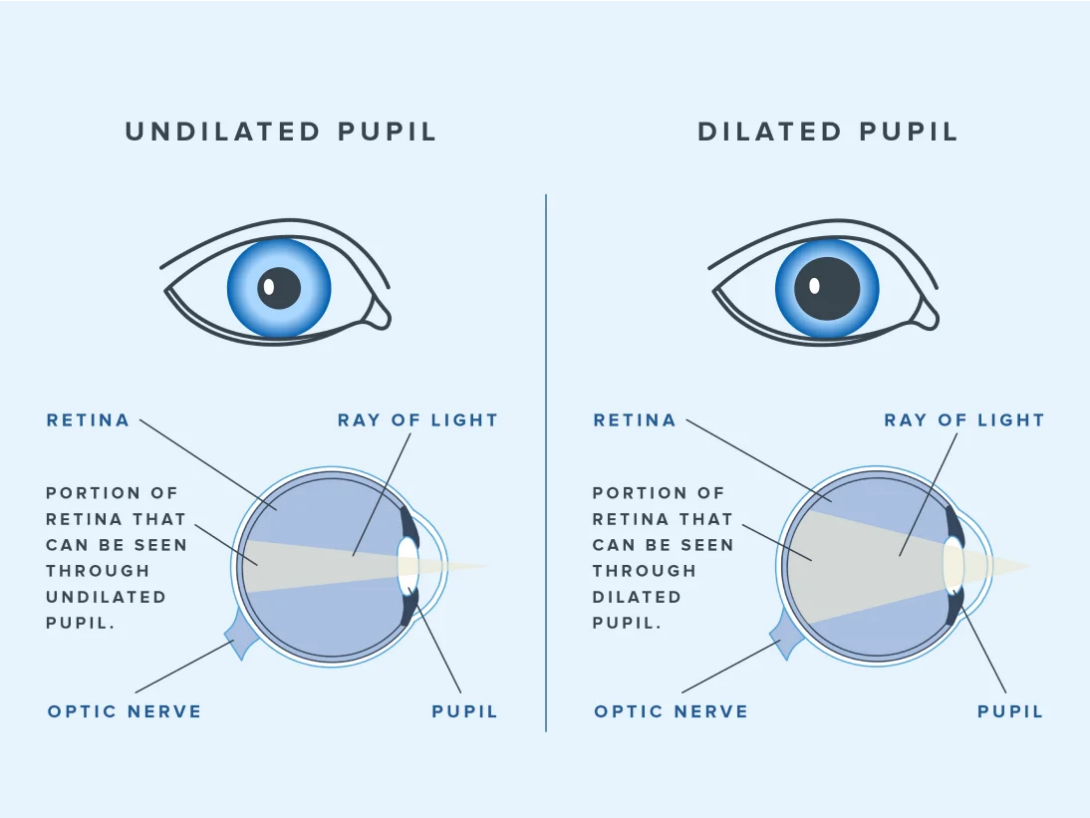

Figure 7: Pupil Dilation

each letter requires a similar amount of horizontal space.

Eye trackers are also less reliable in detecting vertical eye movements. Double-spacing should be used to simplify the process to tell what line of the document is now being read.

Last but not least, when showing larger texts that span numerous displays, the screens must have comparable durations and each stimulus should occur in comparable places.

4. **Regulating non-linguistic visual stimuli**: While presenting visuals, there are several aspects that must be under control. It is crucial to equalize the placement of things on a screen since we typically scan visuals from left to right (for language whose writing is from left to right). If condition y always shows on the left side of the screen and condition x always appears on the right, for instance, condition y would probably always be fixed first—not due to the experimental manipulation, but rather due to its location on the screen. The visuals should also be coordinated for size & salience because it has been discovered that these factors affect gazing patterns.

5. **Take the limitations of eye tracking into account**: Although eye-tracking has indeed been acclaimed as enabling "natural" reading, it does not necessarily mean that we can basically give participants "real" material (like a newspaper story, TOEFL/IELTS reading passage, etc.) and make conclusions directly from various reading time is long for specific words or sentences. It is essential to remember that the readings record may be affected by such factors if experimental material were not thoroughly regulated and prepared, in accordance with the procedures described above. This will cast doubt on any inferences that are taken from the data.

7.3 Hallucination Detection using Gaze Features

Many existing methods for hallucination detection depend on knowledge sources that are explicit such as Wikipedia or knowledge graphs (Manakul et al., 2023; Santhanam et al., 2021; Dziri et al., 2021a; Ji et al., 2023) or ingrained in language encoders such as BERT or RoBERTa (Shen et al., 2023; Zhou et al., 2020). While these traditional approaches can reasonably detect hallucinations in a text when

supplemented with knowledge sources, they face sustainability challenges due to the constant need for up-to-date knowledge. Obtaining the latest information for hallucination detectors is often impractical, as it requires readily available and current sources of knowledge. To address this issue, Maharaj et al. (2023) propose an alternative approach that leverages cognitive and behavioural information from humans in the form of gaze patterns while they analyze text for potential hallucinations.

The work of Maharaj et al. (2023) is motivated by the notion that humans, while reading text for hallucination identification would naturally employ their cognitive faculties to navigate the intricate relationship between language and real-world knowledge. Linguistically, this involves scrutinizing whether (a) entities in the text are adequately placed (e.g., is *Canada* a right choice of entity) (b) the semantic roles played by the entities are valid w.r.t the context (e.g, *Pluto is a planet*). This critical examination would often manifest as prolonged fixations on specific sections of text (e.g., longer fixations on entities and phrases such as Canada, Pluto and a planet) that require closer evaluation, resulting in denser and more extensive fixation activities. In essence, fixations may serve as invaluable indicators, acting as a reliable surrogate for knowledge-based validation of contextual information pertaining to potential hallucinations. This may open up possibilities for the development of a hallucination detector that can leverage gaze data as a primary input, alleviating the reliance on supplementary external knowledge.

To validate this, they first collect a first-of-its-kind eye-tracking data of 5 annotators annotating 500 instances of claim-context pairs, carefully derived from the FactCC dataset (Kryscinski et al., 2020). The annotators were asked to verify whether the claim is consistent with respect to the context or not. During the annotation process, Maharaj et al. (2023) capture the fixation patterns of annotators on both the claim and context texts, along with their corresponding labels. Behavioural analysis of the annotated data reveals a recurrent pattern where annotators tend to skim through somewhat irrelevant context while selectively focusing on information crucial for establishing or refuting hallucinations. Building upon insights gained from this behavioural analysis, they term this selective reading phenomenon as *"attention bias"*. Furthermore, their observations indicate that attention bias can manifest as either a *"global"* approach, involving the extraction of sentences containing relevant information about hallucinations, or a *"local"* approach, focusing on specific phrases within sentences to evaluate the alignment of semantic roles between the claim and the selected phrases.

Building upon these insights, they also propose a modular architecture that incorporates global and local attention bias using transformer-based deep learning techniques Vaswani et al. (2017) and a gaze-based attention saliency module Sood et al. (2020b). Experimental evaluations on the FactCC dataset demonstrate the efficacy of this approach, outperforming baseline models while attaining better interpretability.

8 Future Directions

As research into the phenomenon of hallucination in large language models (LLMs) advances, several pivotal questions warrant sustained exploration and discussion. One critical area of investigation is the efficacy of LLMs' self-correction mechanisms in mitigating hallucinations. Future research should focus on developing and rigorously testing these mechanisms to enhance their ability to identify and correct inaccuracies in real-time, thus improving the overall reliability of generated content.

Another essential avenue for future research is the understanding of knowledge boundaries within LLMs. It is crucial to delineate the limits of what these models know and identify the thresholds at which they are likely to produce hallucinatory information. By mapping these boundaries, researchers can better predict and prevent instances of hallucination, thereby enhancing the practical utility of LLMs in various applications.

Moreover, striking a balance between creativity and truthfulness in LLM outputs remains an open question. While creativity is a valuable attribute that enables LLMs to generate engaging and novel content, it must be tempered with accuracy to avoid the dissemination of false or misleading information. Future studies should investigate methods to fine-tune this balance, ensuring that LLMs maintain their creative potential without compromising factual integrity.

Addressing these questions not only contributes to a deeper understanding of the capabilities and limitations of LLMs but also provides critical insights into the complex nature of hallucinations in machine-generated text. By delving into these

future directions, researchers can pave the way for the development of more robust, accurate, and trustworthy LLMs.

9 Summary & Conclusion

This survey has examined the critical challenge of hallucination detection in large language models (LLMs). We categorized hallucinations into intrinsic and extrinsic types, and further into factuality and faithfulness hallucinations, providing a framework for targeted detection and mitigation. We explored the origins of hallucinations from training data, training phase, and inference, and reviewed diverse detection methodologies including fact overlap-based, entailment-based, classifier-based, question-answering, retrieval-based, uncertainty-based, prompting-based, and gaze-based approaches. Additionally, we discussed key benchmarks like FACTOR, FreshQA, MedHALT, HaluEval, and FELM for performance evaluation. The survey also highlighted the intersection of human cognitive behavior and hallucination detection, introducing terminologies related to eye tracking and discussing how gaze features can be leveraged for this purpose. Understanding human cognitive responses to hallucinations can inform the development of more intuitive and effective detection systems.

In discussing future directions, we highlighted the need to enhance LLMs' self-correction mechanisms, better understand their knowledge boundaries, and balance creativity with truthfulness. These areas are crucial for improving the accuracy, reliability, and practical utility of LLMs. By addressing these challenges, future research can develop more robust and trustworthy LLMs, enhancing their application across various domains.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial nli for factual correctness in text summarisation models. *arXiv preprint arXiv:2005.11739*.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text. *arXiv preprint arXiv:2402.04335*.

Jan Dirk Blom. 2010. *A dictionary of hallucinations*. Springer.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

David Chiang and Peter Cholak. 2022. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Kathy Conklin and Ana Pellicer-Sánchez. 2016. Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3):453–467.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021a. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021b. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*, 4.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Boris A Galitsky. 2023. Truth-o-meter: Collaborating with llm in fighting its hallucinations.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Eckhard H Hess and James M Polt. 1960. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks,

Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. *arXiv preprint arXiv:2104.05433*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3):598–612.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. 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)*, pages 540–551.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Llms as factual reasoners: Insights from existing benchmarks and beyond. *arXiv preprint arXiv:2305.14540*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Zuchao Li, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. 2023b. Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bingbin Liu, Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. 2024. Exposing attention glitches with flip-flop language modeling. *Advances in Neural Information Processing Systems*, 36.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023a. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023b. Chatgpt as a factual inconsistency evaluator for text summarization.

Fiona Macpherson and Dimitris Platchias. 2013. *Hallucination: Philosophy and psychology*. MIT Press.

Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra, and Pushpak Bhattacharyya. 2023. Eyes show the way: Modelling gaze behaviour for hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11424–11438, Singapore. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. 2004. The role of fixational eye movements in visual perception. *Nature reviews neuroscience*, 5(3):229–240.

Ethel Matin. 1974. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020b. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online.

George W McConkie, Paul W Kerr, Michael D Reddix, and David Zola. 1988. Eye movement control during reading: I. the location of initial eye fixations on words. *Vision research*, 28(10):1107–1118.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. If beam search is the answer, what was the question? *arXiv preprint arXiv:2010.02650*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Chau Nguyen, Thanh Tran, Khang Le, Hien Nguyen, Truong Do, Trang Pham, Son T Luu, Trung Vo, and Le-Minh Nguyen. 2024. Pushing the boundaries of legal information processing with integration of large language models. In *JSAI International Symposium on Artificial Intelligence*, pages 167–182. Springer.

OpenAI. 2023. Gpt-4 technical report.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keith Rayner. 1979. Eye guidance in reading: Fixation locations within words. *Perception*, 8(1):21–30.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Michael Bendersky, and Marc Najork. 2023. " why is this misleading?": Detecting news headline hallucinations with explanations. *arXiv preprint arXiv:2302.05852*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Felix Stahlberg, Ilia Kulikov, and Shankar Kumar. 2022. Uncertainty determines the adequacy of the mode and the tractability of decoding in sequence-to-sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8634–8645, Dublin, Ireland. Association for Computational Linguistics.

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

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Yixu Wang, Wenpin Qian, Hong Zhou, Jianfeng Chen, and Kai Tan. 2023b. Exploring new frontiers of deep learning in legal practice: A case study of large language models. *International Journal of Computer Science and Information Technology*, 1(1):131–138.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020b. Towards faithful neural table-to-text generation with content-matching constraints. *arXiv preprint arXiv:2005.00969*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. *arXiv preprint arXiv:2404.04302*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Alexandre Zénon. 2019. Eye pupil signals information gain. *Proceedings of the Royal Society B*, 286(1911):20191593.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.