

# Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art

NEELOY CHAKRABORTY, University of Illinois Urbana-Champaign, USA

MELKIOR ORNIK, University of Illinois Urbana-Champaign, USA

KATHERINE DRIGGS-CAMPBELL, University of Illinois Urbana-Champaign, USA

Autonomous systems are soon to be ubiquitous, spanning manufacturing, agriculture, healthcare, entertainment, and other industries. Most of these systems are developed with modular sub-components for decision-making, planning, and control that may be hand-engineered or learning-based. While these approaches perform well under the situations they were specifically designed for, they can perform especially poorly in out-of-distribution scenarios that will undoubtedly arise at test-time. The rise of foundation models trained on multiple tasks with impressively large datasets has led researchers to believe that these models may provide “common sense” reasoning that existing planners are missing, bridging the gap between algorithm development and deployment. While researchers have shown promising results in deploying foundation models to decision-making tasks, these models are known to hallucinate and generate decisions that may sound reasonable, but are in fact poor. We argue there is a need to step back and simultaneously design systems that can quantify the certainty of a model’s decision, and detect when it may be hallucinating. In this work, we discuss the current use cases of foundation models for decision-making tasks, provide a general definition for hallucinations with examples, discuss existing approaches to hallucination detection and mitigation with a focus on decision problems, present guidelines, and explore areas for further research in this exciting field.

**CCS Concepts:** • Computing methodologies → Artificial intelligence; Machine learning; Natural language generation; Machine learning approaches.

**Additional Key Words and Phrases:** Foundation Models, Decision-Making, Hallucination Detection and Mitigation, Survey

## ACM Reference Format:

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art. *ACM Comput. Surv.* 1, 1, Article 1 (January 2025), 55 pages. <https://doi.org/10.1145/3716846>

## 1 Introduction

A great deal of progress has been made in the last decade and a half with regards to the efficacy and efficiency of models for perception, decision-making, planning, and control [88, 195]. Broadly speaking, approaches to these problems fall under one of two umbrellas: hand-engineered model-based systems and data-driven learning-based models [61]. With some deployment scenario in mind, developers may hand-engineer rules [72] or tune a controller [18] to be tested, or in the case

---

This work is supported by the Office of Naval Research under Grant No.: N00014-23-1-2651.

Authors’ Contact Information: Neeloy Chakraborty, neeloyc2@illinois.edu, University of Illinois Urbana-Champaign, Department of Electrical and Computer Engineering, Coordinated Science Laboratory, Urbana, IL, USA; Melkior Ornik, mornik@illinois.edu, University of Illinois Urbana-Champaign, Department of Aerospace Engineering, Talbot Laboratory, Urbana, IL, USA; Katherine Driggs-Campbell, krdc@illinois.edu, University of Illinois Urbana-Champaign, Department of Electrical and Computer Engineering, Coordinated Science Laboratory, Urbana, IL, USA.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 1557-7341/2025/1-ART1

<https://doi.org/10.1145/3716846>

of learning-based models, collect training data and craft some reward function to fit a model to an objective, given said data [75]. In practice, these methods work particularly well in the scenarios that they were specifically designed and trained for, but may produce undesirable results in previously unseen out-of-distribution cases [221]. Designers may choose to add more rules, re-tune their controller, fine-tune their model to a more representative dataset, fix the reward function to handle edge cases, or even add a detector (which may itself be rule-based or data-driven) at test-time to identify out-of-distribution scenarios before calling on the decision-maker [25, 183, 192]. However, even with these changes, there will always be other situations that designers had not previously considered which will come about during deployment, leading to sub-optimal performance or critical failures. Furthermore, the modifications made to the model may have unforeseen effects at test-time like undesired conflicting rules [57] or catastrophic forgetting of earlier learned skills [95].

Informally, classical methods and data-driven approaches lack some form of *common sense* that humans use to adapt in unfamiliar circumstances [64]. More recently, there has been work towards developing multi-modal large language models, which take inputs in the form of images, videos, audio, and text, to tackle more complex language understanding and reasoning tasks [10, 17]. These models are developed by collecting and cleaning an enormous natural language dataset, pre-training to reconstruct sentences on said dataset, fine-tuning on specific tasks (e.g., question-answering), and applying human-in-the-loop reinforcement learning to produce more reasonable responses [1]. Even though these models are another form of data-driven learning that attempt to maximize the likelihood of generated text conditioned on a given context, researchers have shown that they have the ability to generalize to tasks they have not been trained on, and reason about their decisions. Many researchers are specifically exploring the use of large (visual) language models, L(V)LMs, to fill the knowledge gap found in earlier works [42]. As such, these *foundation* models are being tested in tasks like simulated decision-making [84] and real-world robotics [247] to take the place of perception, planning, and control modules. Even so, foundation models are not without their limitations. Specifically, these models have a tendency to *hallucinate*, i.e., generate decisions or reasoning that sound plausible, but are in fact inaccurate or would result in undesired effects in the world [44, 90, 173]. This phenomenon has led to the beginning of a new research direction that attempts to detect when L(V)LMs hallucinate so as to produce more trustworthy and reliable systems. Before these large black-box systems are applied in safety-critical situations, there need to be methods to detect and mitigate hallucinations. Thus, this survey collects and discusses current hallucination mitigation techniques for foundation models in decision-making tasks, and presents potential research directions.

Existing surveys particularly focus on presenting methods for hallucination detection and mitigation in question-answering (QA) [90, 173, 240, 253] or object detection tasks [119]. For example, Ji et al. [90] summarize metrics and hallucination detection and mitigation methods for abstractive summarization, dialogue generation, generative question-answering, data-to-text generation, and machine translation for natural language generation models in general. Ye et al. [240] and Rawte et al. [173] take this work a step further by classifying hallucination mitigation strategies for foundation models in particular, in generative text and multi-modal settings respectively. Zhang et al. [253] similarly tackle hallucination detection for language foundation models, but limit their definition of hallucinations to conflicts between generations and inputs, contexts, and facts. The authors also discuss other possible problems in generated outputs, like ambiguity and bias. In the domain of image captioning, Li et al. [119] provide examples of hallucinations and present a new metric to evaluate object hallucinations. More recent surveys, like the ones from Bai et al. [10] and Liu et al. [127], dive deeper into hallucination mitigation methods for multi-modal foundation models applied to image captioning tasks, but lack a general definition for hallucinations that encompasses decision-making applications. The extensive survey from Huang et al. [81] primarily

focuses on methods for evaluating the trustworthiness of language models from the fronts of factuality and faithfulness. Their taxonomy also briefly touches on hallucinations in object detection using LVLMs, but like other surveys, ignores broader decision-making deployments. There are also other works that provide examples of current use cases of L(V)LMs in autonomous vehicles [236] and robotics [247, 249]. Wang et al. [212] perform a deep analysis of the trustworthiness of a variety of foundation models and Chen and Shu [27] provide a taxonomy of hallucinations within LLMs, but both exclude applications to general decision problems. To the best of our knowledge, we are the first to propose a general definition of hallucinations that can be flexibly tuned to any particular deployment setting, including commonly found applications to QA or information retrieval, and more recent developments in planning or control. Furthermore, there is no existing work that summarizes state of the art methods for hallucination detection and mitigation approaches within decision-making and planning tasks. Additionally, to the best of our knowledge, ours is the first review to provide guidelines for choosing and designing hallucination intervention algorithms across different application areas.

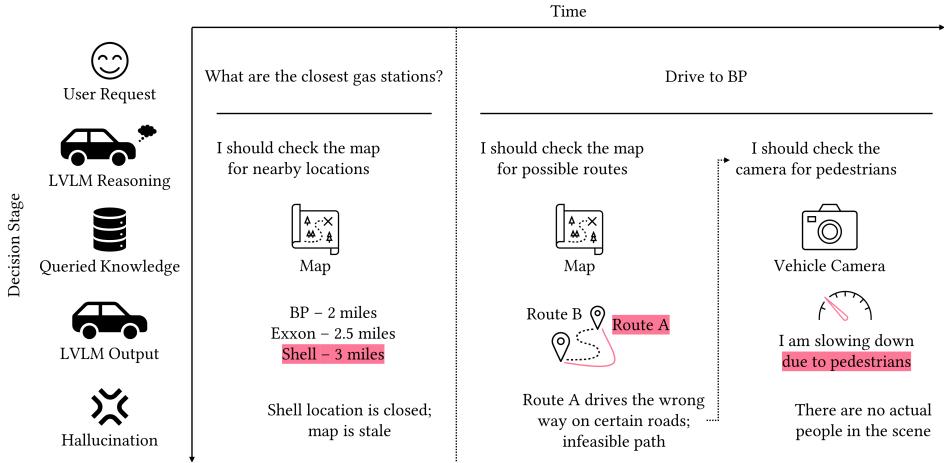
In the remainder of this work, we discuss the current uses of foundation models for decision-making tasks in Section 2, define and provide examples of hallucinations in Section 3, identify current detection methods in Section 4, present guidelines in Section 5, and explore research directions in Section 6.

## 2 Foundation Models Making Decisions

Originally coined by Bommasani et al. [17], the term *foundation models* refers to models that are “trained on broad data at scale such that they can be adapted to a wide range of downstream tasks.” This approach is in contrast to works that design and train models on a smaller subset of data for the purpose of being deployed to a specific task [233]. The key difference is that foundation models undergo a pre-training procedure on a large-scale dataset containing information from a variety of possible deployment fields, through which they are expected to learn more general features and correspondences that may be useful at test-time on a broader set of tasks [256, 258]. Examples of existing pre-trained foundation models span language [20, 48, 206], vision [24, 99, 150], and multi-modal [1, 167] inputs. In this section, we give a brief overview of existing use cases for foundation models in robotics and autonomous vehicles, and we provide a discussion of other decision-making systems in Appendix A.2. We also succinctly point out hallucinations found in these works and leave a lengthier discussion in Section 3.2. Readers should refer to works from Cui et al. [42], Yang et al. [236], Zeng et al. [247], and Zhang et al. [249] for a deeper review of application areas.

### 2.1 Autonomous Driving

For the autonomous vehicle domain, researchers have formulated the use of language foundation models as a fine-tuning and prompt engineering problem [220, 221]. An external sub-system is usually designed with (1) a perception module to process signals from raw sensors, (2) a memory bank of prior important experiences and its corresponding similarity function to find alike scenarios, and (3) a prompt generator to convert current sensor data and relevant memories into natural language that can be input to the foundation model. Currently, works either fine-tune LLMs with a few examples, or directly apply the model in a zero-shot manner, on a QA task with driving related questions. By framing the task in a QA form, researchers have been able to provide context to the L(V)LM to probe for high-level natural language decisions [220, 221], path planning [137, 191], vehicle tracking and trajectory prediction [96, 224], descriptions of the surroundings of the vehicle [29, 231], and low-level control [128]. Figure 1 is an example of how a foundation model may be used in an autonomous driving setting, with possible hallucinations in deployment.



**Fig. 1. Example deployment of an LVLM foundation model in an autonomous driving setting.** Hallucinations (pink) may arise at any point in the decision-making pipeline, including information retrieval, planning, perception, and control. In this example, the LVLM correctly queries the map for possible destinations of gas stations, but lists a location that is no longer open. Then, when navigating to one of the locations, the model predicts a path on the map that goes in the wrong direction of traffic flow. Finally, when applied to perception tasks for detecting possible pedestrians in front of the vehicle, the model hallucinates nearby people, causing an improper control action.

**High-level Decisions.** Wen et al. [220] propose DiLu, a framework consisting of reasoning, reflection, and memory modules that support an LLM in producing high-level decisions for autonomous driving, and they test their method within a driving simulator environment. Specifically, the reasoning module views the current observation of the vehicle, queries the memory module for any similar situations that were encountered in the past, and converts the experience into a prompt, which is input to the LLM. The prompt is formatted such that it elicits chain-of-thought reasoning [217] from the LLM, which is shown to improve the accuracy of the model. The generated text output by the LLM is summarized by the reflection module, and is used to update the memory bank of experiences. A separate decision decoder model converts the summary into a discrete high-level decision (*e.g.*, idle, turn right, accelerate, *etc.*) to take in the simulator. The same authors have also experimented with prompting the LVLM model, GPT-4V [1], in a zero-shot manner to describe the surroundings of the vehicle, take high-level decisions, and explain why it believes it would be a good action to take [221]. They find that GPT-4V is capable of identifying safety-critical scenarios and suggests driving more conservatively in those situations. However, like other existing vision models [198], it has difficulty in detecting traffic light states. We discuss other hallucination examples in traffic scenarios in Section 3.2.

**Path Planning.** Agent-Driver from Mao et al. [137] utilizes a tool library of functions that communicate with neural modules that are responsible for object detection, trajectory prediction, occupancy estimation, and mapping. The LLM is asked to reason about what information would be helpful to plan a path of the ego vehicle, and calls on functions from the tool library to build up relevant context. Like Wen et al. [220], the authors use a memory bank of prior driving experiences to bolster the context provided to the LLM. With this context, the LLM predicts a suitable path for the ego vehicle to follow. If a collision is detected between the predicted trajectory and surrounding

objects in the scene, the LLM undergoes self-reflection, like Reflexion [188], another hallucination mitigation technique, to fine-tune its prediction. Through a short study, the authors test the frequency of invalid, hallucinated outputs from the model, and find that their self-reflection approach results in zero invalid generations at test-time. Sima et al. [191] build up context before predicting a path for the ego vehicle by asking a VLM questions about its perception of surrounding vehicles, predicting their behaviors, planning high-level vehicle decisions, converting to lower level discrete actions, and finally, estimating a coordinate-level trajectory. Their method, DriveLM-Agent, predicts paths from images in an end-to-end manner using a multi-modal approach, whereas Agent-Driver requires sensor modules to process context separately.

*Trajectory Prediction.* Wu et al. [224] propose PromptTrack as a method to predict bounding boxes and trajectories of vehicles in multi-view camera scenes conditioned on a text prompt. PromptTrack is another end-to-end method that encodes multi-view images with an image encoder, decodes previously tracked boxes and current detections into new tracks, uses a language embedding branch to predict 3D bounding boxes, and updates the memory of current tracks using past & future reasoning branches. Rather than using ego vehicle point of view images for object tracking, Keysan et al. [96] propose an approach to convert rasterized images from a birds-eye view of the scene into a prompt describing the past trajectories of vehicles with Bézier curves. This method combines vision and language encoders to generate the Bézier curve-based scene description, and elicits a language model to predict trajectories in a similar format.

*Scene Understanding.* Works using foundation models for generating scene descriptions given multi-modal information frame the task as a QA problem. For example, Chen et al. [29] use a reinforcement learning (RL) agent pre-trained on the driving task in simulation to collect a dataset containing the vehicle’s state, environment observation (assumed to be ground truth from simulator), low-level action, and the ego’s percentage of attention placed on different surrounding agents (if using an attention-based network). The authors ask ChatGPT [147] to act as a professional driving instructor who generates questions and answers given general driving rules and the vectorized description of the scene from the RL agent. Then, combining a pre-trained LLM with a vector embedder and former model, the architecture is trained end-to-end to answer questions from the scene. Examples of questions the architecture is posed at test-time include, “What objects are you observing,” “How are you going to drive in this situation and why,” and, “What are the best tourist spots in London” (the last of which is considered out of scope by the driving model). The authors acknowledge that the LLM may generate undesired hallucinated outputs during deployment, so they augment their training instruction dataset with out of scope questions that the model should learn to refuse to answer. DriveGPT4, proposed by Xu et al. [231], is a multi-modal LLM using LLaMA 2 [207] which encodes video sequences with an encoder model, and projects question and video embeddings into a text form to be input to the LLM. A decoder model converts the tokenized output of the LLM into a low-level action with a corresponding high-level reason to take the action. Like Chen et al., the authors collect a driving QA dataset to align and fine-tune the architecture to driving tasks.

*Control.* Liu et al. [128] tackle low-level control within unsignalized intersections by formulating the problem as a multi-task decision-making process. Specifically, they train expert policies with RL to perform individual control tasks (e.g., turning left, going forward, and turning right), with which they collect an expert demonstration dataset. An LLM model, GPT-2 [168], is fine-tuned to predict next actions given past trajectories of state and action pairs from the expert dataset. The authors showcase that their method, MTD-GPT, is able to achieve higher success rates across each of the

control tasks over the original RL policies alone, showcasing the promise of foundation models to leverage their general knowledge in a multi-task problem setting.

## 2.2 Robotics

Foundation models have also been used in the robotics domain for object detection, affordance prediction, grounding, navigation, and communication. An example of a robot deployed with LVLM capabilities and potential hallucinations is shown in Figure 3 of Appendix A.1. Ichter et al. [86] are motivated by the issue of misalignment between the capabilities of a robot and what an LLM believes it is capable of performing. Because LLMs may not specifically be trained with data from the robot it is to be deployed on, there is a gap in the model’s understanding and the true capacity of the robot, which could lead to hallucinated generations that cannot feasibly be used at runtime. The authors propose SayCan as a method to combine the general knowledge of LLMs with the specific capabilities of a robot in the real-world. Specifically, an LLM is given a task in text form, and is asked to output a list of smaller actions to take in order to complete said task successfully. To constrain the LLM to generate possible actions available to the robot, they assume access to (1) the probability distribution of next tokens to generate from the model, and (2) a set of available skills on the robot, with which they compute the probability of the LLM generating each of the skills next. SayCan greedily selects the action that has the highest product of the next token probability from the LLM and the probability of the action actually successfully being executed in the environment, until the model predicts it has completed the task.

Rather than relying purely on textual context, PaLM-E, proposed by Driess et al. [55], is a multi-modal model that converts various sensor inputs (*e.g.*, images) to a token-space embedding that is combined with instruction embeddings to be input to a PaLM LLM [35]. PaLM is used to either answer questions about the surroundings of the robot, or to plan a sequence of actions to perform to complete a task. Driess et al. further acknowledge that the multi-modality of their PaLM-E architecture leads to increased risk of hallucinations.

Inspired by recent promising findings in using foundation models to generate programs [31], other works deploy foundation models to write low-level code to be run on robots. Liang et al. [121] present Code as Policies, which uses LLMs to hierarchically generate interactive code and functions that can be called. As the model writes main code to be run on a robot given an instructive prompt of the task from the user, it identifies functions to call within the higher level code to complete the task successfully. The authors show that LLMs can leverage third party libraries for existing functions, or develop their own library of functions dynamically with custom methods for the task. While the functionality of Code as Policies can be tested easily for low-level skill definitions, longer multi-step problems require testing whether all requested conditions have been met by running the generated code on the robot. As such, Hu et al. [80] propose the RoboEval performance benchmark for testing robot-agnostic LLM-generated code. Specifically, the CodeBotler platform provides an LLM access to abstract functions like “pick,” “place,” and “get\_current\_location” that have the same external interface regardless of the robot to be deployed on. Like Code as Policies, CodeBotler is provided a text instruction from the user and generates code to be tested. Then the RoboEval benchmark uses *RoboEval Temporal Logic* (RTL) to test whether the generated code meets task and temporal ordering constraints provided in the original prompt. Furthermore, they test the robustness of the LLM by passing in several paraphrased prompts to check for consistency across inputs. We discuss similar consistency-checking strategies for identifying hallucinations in decision-making tasks further in Section 4.3.1.

In the space of robot navigation, LM-Nav leverages a VLM and attempts to predict a sequence of waypoints for a robot to follow and visit landmarks described within a language command [187]. Here, the authors use in-context learning [53] to teach GPT-3 [20] to extract desired landmarks

from a natural language instruction. Assuming there are images of the possible landmarks the robot can navigate to in its environment, LM-Nav uses CLIP [167] to predict the closest matching pairs of extracted landmark descriptions and waypoint images. Finally, dynamic programming is applied on the complete graph of the environment to optimize the path of landmarks to visit. The overall predicted path is optimized to maximize the likelihood of successfully completing the instruction input to the model.

### 3 Hallucinations

Even with all their success on a multitude of deployment areas, foundation models still produce inconsistent outputs, or *hallucinate*, at test-time. Here, we provide a general definition for hallucinations that can be applied to any foundation model deployment task, including various autonomous systems. Additionally, we give examples of hallucinations encountered in literature, and discuss how they come about during testing.

#### 3.1 What are hallucinations?

Across current literature on foundation models, there exist similar patterns and themes that can be used to develop a unified definition for hallucinations. With the majority of works studying this problem within QA tasks, where ground truth answers are available, several authors explain hallucinations as producing text that includes details/facts/claims that are fictional/misleading/fabricated rather than truthful or reliable [173]. Works making use of a dedicated knowledge-base further describe hallucinations as generating nonsensical or false claims that are unsubstantiated or incorrectly cited [26, 117, 144, 250]. Varshney et al. [210] also present the idea that foundation models may sound syntactically correct, or *coherent*, while simultaneously being incorrect. Gallifant et al. [66], who perform a peer review of the GPT-4 technical paper, state that hallucinations include responses that are irrelevant to the original prompt. Li et al. [119], who specifically explore hallucinations of LVLMs in detecting and classifying objects within images, define hallucinations as generating object descriptions inconsistent with target images. A common theme among existing hallucination definitions for QA, information retrieval, and image captioning domains is that, while the generation may sound coherent, either the output is incorrect, or the model’s reasoning behind the generated text is incorrect. However, we find these characteristics on their own do not completely encompass the hallucinations found in decision-making tasks in literature, thus requiring additional nuances.

Within papers that apply foundation models to decision-making tasks specifically, researchers have encountered similar problems of hallucinations impacting performance. Park et al. [154] describe hallucinations as predicting an incorrect feasibility of an autonomous system when generating an explanation behind the uncertainty of an action to take. Similarly, Kwon et al. [104] find that language models may provide incoherent reasoning behind their actions. Wang et al. [215] and Ren et al. [174] believe that these generative models also have a sense of high (false) confidence when generating incorrect or unreasonable plans. In the case of robot navigation and object manipulation, Hu et al. [80] and Liang et al. [122] refer to hallucinations as attempting to interact with non-existent locations or objects.

In the code generation task, Chen et al. [31] use the term “alignment failure,” with similar effects to those of hallucinations discussed above. More specifically, the authors informally describe an alignment failure as an outcome where a model is *capable* of performing a task, but *chooses* not to. If a model is able to complete a task successfully within its latent space (perhaps through additional prompt engineering or fine-tuning), one may ask, “Why would the model *choose* not to?” As foundation models are trained with the next-token reconstruction objective on a training set, they attempt to maximize the likelihood of the next token appearing at test-time as well. Consequently,

**Table 1. Definitions for compliance, desirability, relevancy, and plausibility – the four characteristics of hallucinations.**

Characteristic	Definition
Compliance	Generation meets hard constraints that cannot be ignored
Desirability	Generation attempts to meet soft constraints measured by some cost or reward function
Relevancy	Contents of generation do not fall outside of a defined set of topics
Plausibility	The syntactic similarity of a generation and in-domain normal samples measured via a critic function

if the test-time prompt includes even minor mistakes, Chen et al. find that LLMs will continue to generate buggy code to match the input prompt. This issue is further described in Section 3.3.

We realize existing definitions for hallucinations are extremely disparate depending on the deployment area. Nevertheless, we have identified three distinct features that are commonly missing within hallucinated generations: *compliance*, *desirability*, and *relevancy*. Additionally, we find hallucinations may seem to be *plausible* while they are in fact unacceptable. A compliance metric checks that the generation meets hard constraints while the desirability metric measures how well the generation meets soft constraints defined by the model engineer. Irrelevant generations refer to predictions that contain details outside of a set of requested topics. Notice that irrelevant predictions can still be considered compliant and desirable depending on how the hard and soft constraints are defined. Plausibility compares the generation’s syntax to those of a set of known, unhallucinated samples. While relevancy is evaluating the content of a prediction, plausibility is assessing its phrasing. Table 1 provides complete definitions for each of the four characteristics.

Then, to bridge definitions from existing QA application areas, decision-making tasks, and all other possible test scenarios for foundation models, we combine these findings and define the term hallucination as follows:

**Definition 3.1.** *A hallucination is a generated output from a model that conflicts with constraints or deviates from desired behavior in actual deployment, or is completely irrelevant to the task at hand, but could be deemed syntactically plausible under the circumstances.*

There are three key pieces to this definition:

- (1) A generated output from a model.
- (2) A deployment scenario to evaluate model outputs with any of the following:
  - A list of constraints that must be *compliant* within the generation.
  - A loose interpretation of a *desired behavior* the generation should meet.
  - A set of topics *relevant* to the task.
- (3) Metrics measuring compliance, desirability, relevancy, and syntactic soundness (*plausibility*) of generations.

In practice, this definition generally encapsulates the qualities of hallucinations discussed earlier. For example, in QA or image captioning tasks, one may define a set of relevant topics that a generation should not stray from, and constraints may be held in the form a knowledge-base of ground truth facts. The desired behavior of the generation may be to be phrased in an informative manner, rather than sarcastic. On the other hand, in robot manipulation settings, a developer may have a set of constrained actions feasible on the robot, and the desired behavior could be to complete a task with as few actions as possible. Relevancy may be measured in relation to the specific task to be deployed on (e.g., a prompt requesting a recipe to make pasta would find it irrelevant if the model also suggested a song to play while cooking). Finally, plausibility informally relates to a measure of how believable an output is to a critic. A more realistic generation has a greater

**Table 2. Examples of applying Definition 3.1 to different tasks.** Note that developers may choose to only define a subset of hallucination characteristics for their deployment depending on evaluation preferences. The table is split into non-decision-making and decision-making applications.

Problem Setting	Characteristic			
	Compliance	Desired Behavior	Relevancy	Plausibility
Question-Answering	Generations must align with database facts	Tone of answer should be informative	Answers should not include references to unrelated topics	Generation is syntactically sound and believable
Image Captioning	Objects in description must appear in image	Censor descriptions for inappropriate images	Descriptions should not be embellished with details that cannot be confirmed	
Planning	Predicted sub-task must be feasible to solve	Plans should maximize expected return	Predicted sub-tasks and actions should not stray from the end goal with added steps	Generated plan is reasonable and seems to attempt to accomplish goal
Control	Predicted action must be possible to perform	Predict actions to complete plan efficiently		

chance of deceiving the user into trusting the model, even when the plan may be hallucinated. Overall, hallucinated outputs may contain one or more of the core characteristics (noncompliant, undesired, irrelevant, and plausible) simultaneously, and our definition can be flexibly applied to any deployment scenario in mind by choosing metrics for each characteristic, respectively. We show more examples of applying our definition to various tasks in Table 2.

### 3.2 Examples

*Driving Tasks.* As discussed in Section 2.1, Wen et al. [221] test GPT-4V on the autonomous driving task and identify failure modes. Regardless of the weather and driving conditions, GPT-4V has difficulty detecting and identifying the traffic light state at an intersection, until the image has zoomed in on the light itself. It also presents additional irrelevant (or completely false) details about other agents, when the prompt had no mention of them in the first place. Furthermore, the model also has difficulty in describing temporal sequences (*i.e.*, videos) and categorizing images by their direction within a panoramic view from the vehicle’s perspective. In their later work, Wen et al. [220] describe that hallucinations arise in these complex environments because of the high variability in driving scenarios. Even after applying hallucination mitigation techniques like chain-of-thought reasoning, the model is not free of these undesired outputs. A similar work evaluating the frequency at which LLMs hallucinate in their descriptions of images, finds that these models’ outputs may include non-existent objects, or additional irrelevant phrases (that may not even be possible to test for accuracy) [119]. For example, in a picture of food on a table, an LLM hallucinates a non-existent beverage, and predicts that the “table is neatly arranged, showcasing the different food items in an appetizing manner.” Although the classification error and irrelevant generation in this example are not critical, earlier works warn of possible failures with more severe, high societal impact (*e.g.*, biases in models leading to marginalizing users) [17].

*Code Generation.* Chen et al. [31] explore alignment failures of LLMs applied to code completion tasks. The authors evaluate the likelihood of these models generating defective code given different input prompts, and discover that in-context learning using examples with buggy code has a higher chance of resulting in poor generations from the model on the actual task at hand. The study also identifies similar model biases towards race, gender, religion, and other representations. Furthermore, the authors find that their model, Codex, is able to generate code that could assist with developing insecure applications or malware, albeit in a limited manner. These findings have been corroborated by other foundation model code generation works in the robotics domain. For example, Wang et al. [213] describe that Voyager sometimes generates code with references to items that do not exist within MineDojo. Similarly, Hu et al. [80] find that their model has the tendency to

call functions with invalid objects or locations, pickup objects when it is already holding something, ask for help when no one is near, and other undesired behaviors.

*Question-answering Domain.* Several works focus on identifying cases of hallucinations in QA tasks. Although this application area is not the direct focus of this work, we present examples of hallucinations in this field as we can glean similar failure modes that could arise within decision-making systems. Common hallucinations in QA result in incorrect answers to questions. For example, Achiam et al. [1] find that GPT-4 “hallucinates facts and makes reasoning errors.” Achiam et al. categorize these failures into closed-domain (given context, the model generates irrelevant information that was not in the context) and open-domain (the model outputs incorrect claims without any context) hallucinations. After fine-tuning on more data with a hallucination mitigation objective, the model reduces its tendency to hallucinate, but still does not achieve perfect accuracy – a similar trend encountered by Touvron et al. [206]. Another set of works identify hallucinations with contradictions among several sampled generations from an LLM, discussed further in Section 4.3.1 [144, 250]. Intuitively, if a context passed into a model results in conflicting generations, the model must be hallucinating some part of the output. Notice in this example, with relation to Definition 3.1, self-contradiction works test for compliance by checking *consistency* among multiple (hallucinated) generations, rather than with respect to a ground-truth knowledge-base that usually exists in QA tasks. As such, our definition can flexibly apply to different system setups by describing compliance, desired behavior, and relevancy respectively.

Additional examples of hallucinations encountered in image, video, and 3D generation methods, and broader impacts faced by medical, legal, and finance industries are discussed in Appendix B.1 and B.2.

### 3.3 Why do they happen?

There are several speculations as to how hallucinations come about during deployment. First and foremost, like any learning task, foundation models are sensitive to biases in training data [173]. Once a model is trained on a given large dataset, some facts may become out-of-date or stale at any point in time [162]. Furthermore, as the training set is embedded into a smaller encoding dimension, the knowledge within an L(V)LM’s frozen parameters is lossy, and models cannot feasibly be fine-tuned every time there is new data [58, 157]. Zhang et al. [250] recommend changing algorithm parameters at runtime, such as, *temperature* (spread of probability distribution of next token), *top-K sampling* (narrows the set of next tokens to be considered), and *beam search* (choosing a set of possible beams, i.e., trajectories, of next tokens based on high conditional probabilities), but the process of tuning these parameters is expensive.

To combat out-of-date training data, some works provide models with an external knowledge-base of information to pull facts from, with the hope of increasing model accuracy. Even with this up-to-date information, Zhang et al. [251] pose that there may exist a misalignment between the true capabilities of a model, and what a user believes the model is capable of, leading to poor prompt engineering. In fact, poor prompting is one of the most significant causes of hallucinations. Chen et al. [31] find that poor quality prompts lead to poor quality generations, in the context of code completion. This phenomenon is attributed to the reconstruction training objective of LLMs attempting to maximize the likelihood of next generated tokens, given context and past outputs [141], i.e.,

$$\log p(s|x) = \sum_{i=1}^N \log p(\sigma_i|x, \sigma_{i-k} \dots \sigma_{i-1})$$

where  $x$  is a context input to the model,  $s$  is an output sequence of  $N$  tokens  $\sigma_1 \dots \sigma_N$ , and any generated token  $\sigma_i$  is conditioned on  $k$  previously generated tokens. As the public datasets these models are trained on contain some fraction of undesirable generations (e.g., defective code), the models become biased to generate similar results under those inputs. Qiu et al. [164] show that this limitation can actually be exploited to push foundation models to generate toxic sentences, or completely lie, by simply rewording the prompt.

While foundation models condition generated tokens on ground-truth text without hallucinations at train time, during inference, the model chooses future tokens conditioned on previously (possibly hallucinated) generated text. As such, Chen et al. [33] and Varshney et al. [210] state that generated outputs are more likely to contain hallucinations if prior tokens are hallucinated as well. Furthermore, Li et al. [115] find that, even if prompt context provided to a foundation model is relevant, the model may choose to ignore the information and revert to its own (possibly outdated or biased) parameterized knowledge.

Overall, the hallucination detection task is highly complex with several possible sources of failures that need to be considered at test-time. Chen and Shu [27] validate the complexity of the detection problem with studies identifying that human- and machine-based detectors have higher difficulty correctly classifying misinformation generated from LLMs than those written by other people.

## 4 Detection and Mitigation Strategies

Hallucination detection and mitigation methods can be classified into three types (white-, grey-, and black-box) depending on the available inputs to the algorithm. Generally, given some context, a foundation model outputs a predicted sequence of tokens, the corresponding probabilities of each token, and embeddings of the generation from intermediate layers in the network. White-box hallucination detection methods assume access to all three output types, grey-box require token probabilities, and black-box only need the predicted sequence of tokens. Because not all foundation models provide access to their hidden states, or even the output probability distribution of tokens (e.g., the ChatGPT web interface), black-box algorithms are more flexible during testing. In this section, we present existing detection and mitigation approaches clustered by input type. While several of these works show promise in QA and image captioning settings, many of them require further validation on decision-making tasks, and we will point out these methods as they come about. Works in this section are summarized in Table 3. We also provide additional details about the evaluation setups of works deployed on custom datasets, simulators, or the real-world. Certain methods that are less related to decision-making are described in Appendix C. Additionally, frequently used metrics, datasets, and simulators are summarized in Appendix D.

### 4.1 White-box Methods

Methods in this section require access to internal weights of the model for hallucination detection.

**4.1.1 Hidden States.** Some approaches utilize intermediate embeddings at different network layers.

- (1) Azaria and Mitchell [7] empirically find that language models attempt to correct themselves after outputting an untruthful claim. As such, they hypothesize that the internal states of a model must have some understanding of whether the output is correct. Furthermore, the authors stray away from directly using the token probabilities, even though they have correlation with model accuracy [210], because the complete output sentence's probability is dependent on the length of the generation and appearance frequency of tokens. Azaria and Mitchell present SAPLMA, a simple classifier trained with supervised learning, that takes the activation values of a hidden layer of an LLM as input, and outputs the probability

**Table 3. A summary of hallucination detection & mitigation methods discussed in Section 4.** Deployment scenarios are split into question-answering (QA), information retrieval (IR), image captioning (IC), image generation (IG), & planning (P) tasks. The method ID includes the subsection the method appears in the paper and the order in which it appears in the subsection. Bolded method IDs are deployed to decision-making tasks specifically. Custom datasets, custom simulators, & real-world experiments for testing are abbreviated as CD, CS, & RW, respectively.

Modality	Method Type	Method ID	Application		Deployment	Evaluation Setting
			Detection	Mitigation		
White-box	Hidden States	4.1.1.1	•		QA	CD
		4.1.1.2	•		QA	CD
		4.1.1.3	•		QA	DecodingTrust [212] TruthfulQA [125]
	Attention Weights	4.1.2.1	•	•	IC	MSCOCO [126] Visual Genome [100]
		4.1.3.1	•		QA	CD
	Honesty Alignment	4.1.3.2	•	•	QA	TriviaQA [92]
		4.2.1.1	•	•	IR/QA	HotpotQA [237] CD
Grey-box	Concept Probabilities	4.2.1.2		•	IC	MSCOCO [126]
		4.2.1.3		•	P	Ravens [246] BabyAI [34] VirtualHome [161]
		4.2.2.1	•	•	IR/QA	MIMIC-CXR [91] CNN/DM [77] TriviaQA [92]
	Conformal Prediction	4.2.2.2	•	•	QA	MMLU [76]
		4.2.2.3	•	•	P	TableSim [174] RW
		4.2.2.4	•	•	P	TableSim [174] CD
	4.2.2.5		•	•	P	CS
Black-box	Analyzing Samples	4.3.1.1	•		IR/QA	CD
		4.3.1.2	•	•	IR/QA	CD
		4.3.1.3	•	•	IR/QA	GSM8K [38] MMLU [76] CD
	4.3.1.4				P	SaGC [154] CS RW

Continued on next page

Table 3 – continued from previous page

Modality	Method Type	Method ID	Application		Deployment	Evaluation Setting
			Detection	Mitigation		
Black-box	Analyzing Samples	4.3.1.5	•	•	IR/QA	CD
		4.3.1.6	•	•	IR/QA	Quest [134] MultiSpanQA [116] FActScore [142]
	Adversarial Prompting	4.3.1.7	•		IC/QA	MSCOCO [126] A-OKVQA [185] GQA [85]
		4.3.1.8	•		QA	BIG-Bench [196] GSM8K [38] MMLU [76]
		4.3.1.9	•		QA	CD
	Proxy Model	4.3.2.1	•		IG	CD
		4.3.2.2	•	•	IR/QA	Natural Questions [103] CD
	Grounding Knowledge	4.3.2.3	•		QA	CD
		4.3.2.4	•		QA	CD
	Constraint Satisfaction	4.3.3.1	•		IR/QA	CD
		4.3.3.2	•		IR/QA	SQuAD [169] HotpotQA [237] TriviaQA [92]
		4.3.3.3	•		IR/QA	CD
		4.3.3.4	•		IG	CD
		4.3.4.1		•	IR/QA	Natural Questions [103] StrategyQA [68] QReCC [5]
		4.3.4.2	•	•	IR/QA	LC-QuAD [208] KQA-Pro [23] ScienceQA [132]
		4.3.4.3	•	•	IR/QA	FuzzyQA [251]
		4.3.4.4	•	•	IR/QA	CD
		4.3.4.5		•	P	TextWorld [40]
		4.3.4.6	•	•	IG	CD
		4.3.4.7		•	IG	I2P [182]
	4.3.5.1	•	•		P	CD
	4.3.5.2	•			P	RoboEval [80]

of the generated claim being true. The authors choose to collect a new dataset of simple statements with corresponding true/false answers, since popular datasets like FEVER [203] do not partition statements by topic, and some statements cannot be cleanly classified. In total, their dataset contains 6K sentences spanning six topics. SAPLMA is shown to be able to identify untruthful outputs, even when trained on a held-out dataset on a completely different topic than evaluated on.

- (2) Yao et al. [238] aim to test the resiliency of foundation models to varying prompts (a form of adversarial prompting further discussed in Section 4.3.2). They propose perturbing an input prompt with additional tokens so as to make an LLM under test produce a desired hallucination (e.g., modify the original query, “Who won the 2020 US election,” to get the LLM to generate, “Donald Trump was the victor,” while its original response was correctly stated as, “Joe Biden was the victor”). As the search space of possible tokens to add/replace when developing the adversarial prompt is massive, the work uses a gradient-based token replacing strategy. Specifically, they define an objective that attempts to find trigger tokens in the direction of the gradient that maximizes the likelihood of the model outputting the desired hallucination. To evaluate their approach, Yao et al. collect a dataset of truthful facts by feeding questions from Wikipedia [62] into an LLM. Then, certain subjects, objects, or predicates in the generated answers are replaced to manually create hallucinations. With simple prompt modifications, the authors show that the white-box approach is able to induce the specified hallucinations.
- (3) The work from Song et al. [194] is described in Appendix C.1.1.3.

**4.1.2 Attention Weights.** Attention weight matrices, which are prominent within transformer model architectures, signify the importance the model places on earlier tokens within a generation when predicting future tokens.

- (1) OPERA, proposed by Huang et al. [82], is a hallucination detection method for LVLMs that makes use of the model’s internal attention weights. When visualizing the attention matrix, the authors find that there exist peculiar column patterns that align with the beginning of a hallucinated phrase. These *aggregation patterns* usually occur on a non-substantial token like a period or quotation mark, but are deemed to have a large impact on the prediction of future tokens. As such, this finding led Huang et al. to modify the beam search algorithm [63] by applying a penalty term to beams wherever an aggregation pattern is detected, and roll back the search to before the pattern arises. Their method is shown to reduce hallucinations, and even eliminate possible repetitions in generations.

**4.1.3 Honesty Alignment.** In addition to methods that require hidden states or attention matrices, we also include methods that fine-tune foundation models to better communicate their uncertainty to questions under white-box algorithms, as they require access to model weights for training.

- (1) For example, Lin et al. [124] collect a calibration dataset of questions and answers from GPT-3 under 21 types of arithmetic tasks (e.g., add/subtract and multiply/divide), and record how often each task is incorrectly answered. They aim to fine-tune the LLM to also output its certainty that the prediction is correct. Consequently, Lin et al. fine-tune the model with data pairs of a question and the empirical accuracy on the task that the question originates from in the calibration dataset, such that the model is expected to similarly output a probability of accuracy at test-time. The authors show that the proposed verbalized probability in deployment does correlate with actual accuracy on the tasks. Specifically, Lin et al. find that certain problems, like multiplication, are more challenging for GPT-3 to answer correctly. By calibrating the model on a set of simpler questions (add/subtract), the model is able to

generalize its verbal uncertainty to more challenging tasks with multiple possible answer choices.

- (2) The work from Yang et al. [234] is described in Appendix C.1.3.2.

## 4.2 Grey-box Methods

Grey-box approaches leverage the probability distributions of tokens output from the model.

### 4.2.1 Concept Probabilities.

- (1) Empirically, Varshney et al. [210] show that there is a negative correlation between hallucination rate and token probability (*i.e.*, as a token’s probability decreases within a sentence, the tendency to hallucinate increases). Thus, the authors rely on token probabilities to estimate uncertainty of concepts within a generated claim, and they check for correctness by cross-referencing a knowledge-base. Whenever a concept is found to be conflicting with a fact through verification questions, their method attempts to mitigate the error by prompting the LLM to replace the incorrect claim with the evidence. The authors query GPT-3.5 for summaries about 150 different topics sampled from popular Wikipedia subjects, including sports, music, and politics. To ensure the accuracy of labels, the authors manually label the truthfulness of the first five sentences in each generated article, rather than relying on crowd-sourced labels. Although effective in the QA setting, Varshney et al. concede that, in the event token probabilities are not available, some form of heuristic must be used to detect hallucination candidates.
- (2) Zhou et al. [259] show that external models can be developed to automatically *clean* hallucinations. The authors tackle the issue of object hallucinations that LVLMs experience when describing the content of images. Through theoretical formulations, the authors show that LVLM responses tend to hallucinate in three settings: when described object classes appear frequently within a description, when a token output has low probability, and when an object appears closer to the end of the response. As such, their model, LURE, is a fine-tuned LVLM trained on a denoising objective with a training dataset that is augmented to include objects that appear frequently within responses, and replacing objects with low token probabilities or appearing close to the end of the response, with a placeholder tag. At inference time, tokens are augmented similarly to how they were changed to generate the training dataset, and the LURE LVLM is prompted to denoise hallucinations by filling in uncertain objects.
- (3) SayCanPay, proposed by Hazra et al. [73], builds off of the SayCan framework [86] to improve the expected payoff of following a plan specified by a language model. Within our hallucination definition, this goal translates to increasing the desirability of generations by improving the likelihood of the model achieving higher rewards. The authors propose three different strategies for planning: Say, SayCan, and SayCanPay. Say methods greedily choose next actions based only on token probabilities. SayCan approaches also take the success rate of the chosen action into consideration. Finally, SayCanPay additionally estimates the expected payoff from following the plan with some heuristic. Hazra et al. learn this Pay model with regression on an expert trajectory dataset. Combining all three models together minimizes the likelihood that a generated plan contains conflicting infeasible action calls, while maximizing the efficiency of the task completion.

- 4.2.2 *Conformal Prediction.* Another range of works estimate the uncertainty of a model output with conformal prediction so as to provide statistical guarantees on the likelihood of predictions being correct [186].

- (1) Quach et al. [165] propose conformal language modeling to build a set of possible candidate responses to a test prompt, while calibrating algorithm parameters on a held-out dataset of independent prompts and their corresponding admission functions, which check whether a model output meets the criteria of an input prompt. In their algorithm, the authors calibrate thresholds for three separate scoring functions that test for generation quality, similarity with other responses, and model confidence using “Learn then Test” [6]. At inference time, given scoring functions, calibrated thresholds, and an input prompt, the method samples outputs from the model and adds them to a prediction set if they meet quality and similarity thresholds, until the whole set is guaranteed to meet the user-defined confidence parameter.
- (2) The work from Kumar et al. [102] is described in Appendix C.2.2.2.
- (3) While the previous hallucination mitigation works presented using conformal prediction are solely applied to QA settings, Ren et al. [174] are the first to apply conformal prediction of foundation models to robotic tasks. The authors are motivated by a desire for language-conditioned robots to understand when they are uncertain about the next action to take, such that they can ask for help in those cases (while minimizing frequency of clarifications). Because LLM generations with different length sequences inherently produce different complete sentence probabilities, the authors propose framing the control task as a multiple-choice problem, like Kumar et al. [102]. Their approach, KnowNo, prompts an LLM to generate a possible set of next actions to take in multiple choice form. They first collect and hand-label a calibration dataset of pairs of held-out instructions and the probability of the model choosing the best action to take next. At test-time, the model outputs a set of actions to take and KnowNo eliminates actions with token probabilities less than a calibrated certainty from a user. If there are still multiple actions left in the prediction set after eliminating uncertain actions, the model queries a human for help choosing the next action. Ren et al. show that KnowNo deviates from the user-defined error rate least often compared to methods that do not use conformal prediction and has the highest success-to-clarification ratio. Additionally, the authors deploy KnowNo to a real UR5 robot arm, where it is tasked with sorting foods on a table by order of user preference, and disposing of undesired objects conditioned on ambiguous instructions. However, several assumptions had to be made to produce the demonstrated results, including having access to next token probabilities, having resources to collect a large calibration dataset, presuming people will faithfully provide help when asked for it, and using ground-truth vision to fully ground the environmental objects with the text input to the model.
- (4) Liang et al. [122] extend the KnowNo methodology by incorporating an introspective planning step using a previously constructed knowledge-base of experiences, which tends to (1) enhance quality of generated plans, and (2) improve interpretability of decisions. Specifically, introspective planning first constructs a knowledge-base containing training pairs of tasks, observations, and valid plans, which the LLM is prompted to generate explanations behind why they are reasonable. Each experience is stored with a key as an embedding of the original instruction. During inference, given a new test instruction, their method queries the database to find the key with the closest embedding to that of the new instruction. This previous experience and reasoning is fed into the model to generate a set of candidate plans to follow. Finally, the remainder of the algorithm follows the same process as KnowNo to calibrate and narrow down the prediction set to fall within a desired error rate. Liang et al. evaluate their method on the KnowNo dataset [174] and an augmentation of the original dataset that considers more safety-critical tasks with ambiguous instructions (*e.g.*, making sure not to place metal objects in a microwave when tasked with heating a bowl).

- (5) Wang et al. [215] aim to provide additional guarantees on completing the task provided within the natural language instruction. To do so, the authors propose a novel task specification, LTL-NL, which combines linear-temporal-logic (LTL) descriptions with natural language from a user instruction, which the authors claim is easier to define than classical LTL specifications. Given this specification, a symbolic task planner chooses a sub-task to complete next and an LLM generates plans for each sub-task, respectively. Like Ren et al. [174] and Liang et al. [122], Wang et al. apply conformal prediction to minimize the number possible actions to take next within some desired error rate. However, rather than directly asking a user for assistance when there is high uncertainty in the next action to take (or when there are environmental constraints), their method, HERACLES, samples a new sub-task to complete from the task planner. If on the other hand, the task planner is unable to provide a new sub-task, HERACLES requests help from the user. The authors deploy the model and baselines in a custom-made 3D mobile manipulator simulator that allows for evaluation of methods that use LTL specifications. For example, the instruction, “Deliver Apple to A” is converted to LTL specifications and fed into HERACLES to navigate the robot to (1) pick up the requested apple and (2) drop it off at location A. With experimentation, the authors find that their method achieves higher task completion rate on missions requiring more sub-tasks, outperforming baseline planners that do not utilize LTL specifications.

### 4.3 Black-box Methods

Black-box algorithms only rely on the input prompts and output predictions from the model, without making assumptions on the availability of the hidden state, nor the token probabilities.

**4.3.1 Analyzing Samples from Model.** As a result, several works of this type sample multiple sentences from an LLM, and measure the similarity of the information present in all samples.

- (1) For example, SelfCheckGPT, proposed by Manakul et al. [136], samples multiple responses (each of which may contain many sentences) from an LLM to a single query, and measures the consistency among the varied responses through a study with five different approaches. SelfCheckGPT with BERTScore [252], for example, computes a similarity score between two sentences from different sampled outputs. Intuitively, a hallucination is detected when the similarity score for a sentence is low across all other samples. Other consistency-checking methods the authors consider include using an automatic multiple-choice QA system conditioned on the responses, relying on a proxy-LLM which has access to token probabilities, training an external classifier to predict contradictions (coined SelfCheck-NLI), and directly prompting the LLM to evaluate whether any sentence can be supported by another sampled response. Before evaluating the proposed methods, the authors find that there is no standardized hallucination detection dataset for question-answering. Furthermore, the existing hallucination dataset from Liu et al. [130] is generated by manually replacing tokens in actual facts, which may not represent generations from real language models. Thus, Manakul et al. choose to curate their own evaluation dataset by querying GPT-3 for articles on topics from the WikiBio dataset [107]. They then manually label the accuracy of each sentence in each generated article for a total of 1908 sentences across 238 summaries. As expected, Manakul et al. find that sampling more responses from the model leads to better estimation of the validity of a claim, but is slower to compute.
- (2) The work from Elaraby et al. [58] is described in Appendix C.3.1.2.
- (3) Rather than analyzing the responses of a single language model, Du et al. [56] take an ensemble approach. Specifically, they propose pitting multiple instances of an LLM into a debate of the correct answer to a question. In practice, several agents are given the same

question and predict a response. Over multiple iterations, all the responses of other agents are concatenated, and fed in as additional context to each model, and a new response is sampled. Du et al. first identify whether the proposed debate method can improve the *reasoning* of language models through three offline tasks with increasing difficulty in a custom benchmark: (1) simple arithmetic, (2) grade-school math [76], and (3) predicting the next best move to take in a game of chess. They additionally evaluate the *accuracy* of debate results in (1) generating biographies of famous computer scientists from Wikipedia, (2) general exam knowledge [76], and (3) confirming the validity of next moves in chess. The authors show that a combination of their iterative ensemble approach with chain-of-thought reasoning mitigates individual hallucinations (as agents tend to converge on a single consensus) and increases QA accuracy. While only tested on offline datasets, the approach could be utilized within a simulation framework, like the one presented by Park et al. [155], where, for example, multiple language agents may debate about plans to make. It is important to note that Du et al. evaluate the accuracy of biography generations by querying ChatGPT for the consistency between a fact and generated response. We hypothesize that, like Yu et al. [243], this automatic labeling scheme will result in lower quality labels than manual annotation.

- (4) CLARA is a framework engineered by Park et al. [154] that predicts when an instruction provided to a robotic system controlled by an LLM may be ambiguous or infeasible. Intuitively, if a language model is uncertain about an instruction, it might output diverse (or conflicting) actions to other instructions that hold similar information. Thus, CLARA samples several sets of concepts from the original prompt, randomly orders them to assemble multiple inputs to the model, and passes them as input to the language model under test. The method computes the average similarity of pairs of output actions in an embedding space, over all outputs. Next, to check for infeasibility of the original goal, the foundation model is provided the possible action space of the robot, environmental observation, and goal, and is prompted to output whether the desired task is practical. In the event the model is uncertain from the multi-prompting step, but the goal is feasible, CLARA asks for clarification from the user with reasoning for why it is uncertain. In addition to evaluating CLARA on SaGC, a custom dataset described in Appendix D.2.2, the authors also put their model to the test on a tabletop manipulator robot in simulation and the real world. While the method achieves a reasonable success rate on robotic pick-and-place tasks with real user instructions (where other discussed methods are primarily evaluated in QA settings), there are still failure cases where the model hallucinates during uncertainty reasoning and feasibility prediction.
- (5) Another set of works explicitly identify contradictions among responses, instead of estimating similarity. Naturally, detecting a self-contradiction is guaranteed to reveal an invalid claim. Mündler et al. [144] pose that removing detected conflicting information will increase the validity of a generated response. As such, the authors suggest a solution that finds important concepts within a response to be evaluated, prompts the generation model to generate more information about each of the concepts, and uses a separate analyzer language model to evaluate the consistency of pairs of sentences on the same concept. Any sentences that are found to be conflicting are revised by the analyzer model, before being output to the user. The authors ask four language models to generate 360 summaries for 30 diverse topics from WikiBio [107] and Wikipedia [62] and have three human annotators manually label any instances of self-contradiction, inaccurate sentences, and unverifiable statements, leading to high-quality labels. With respect to Definition 3.1, the annotators are identifying cases of noncompliant and irrelevant generations to decide which statements are hallucinations.
- (6) Rather than relying on another language model to analyze the correctness of predictions, Dhuliawala et al. [49] utilize chain-of-thought reasoning to prompt an LLM to generate possible

verification questions about its original responses. If there are conflicts in answers to the verification questions, the LLM is prompted to regenerate its output with updated context and conflicting reasoning. One of the first evaluations the authors perform is on a custom set of 56 questions with topics collected from Wikipedia, each of which have multiple correct answers for a total of around 600 ground-truth entities.

- (7) The work from Li et al. [119] is described in Appendix C.3.1.7.
- (8) The work from Xiong et al. [229] is described in Appendix C.3.1.8.
- (9) The work from Yehuda et al. [241] is described in Appendix C.3.1.9.

**4.3.2 Adversarial Prompting.** Works specializing in adversarial prompting attempt to test the robustness of models to varying inputs that may coerce the model into producing out-of-distribution results.

- (1) For example, Mehrabi et al. [138] apply adversarial prompting to text-to-image foundation models, like Stable Diffusion [182], to generate offensive images. With respect to Definition 3.1, their framework, FLIRT, is essentially testing the tendency of foundation models to hallucinate undesired generations in deployment. FLIRT uses an adversarial language model to predict a prompt to input to the image generator, scores the generated image for the presence of undesirable traits using an external classifier, re-prompts the adversary to produce a new instruction conditioned on the findings of the classifier, and repeatedly generates images until the adversary successfully prompts the test model to output an undesirable result. Mehrabi et al. define objective functions conditioned on the score output by external classifiers to maximize diversity of adversarial prompts and minimize toxicity so as to pass text filters that detect malicious inputs, while improving attack effectiveness. The authors form a large set of prompts with varying levels of detail across different sexual, violent, hate, etc. contexts, inspired by Schramowski et al. [182]. Each prompt corresponds to one of three test splits that change the type of toxicity, level of detail, and phrasing in the prompt. Tangentially, Mehrabi et al. evaluate FLIRT on text-to-text models by similarly collecting adversarial prompts with varying vulgarity.
- (2) Another work from Yu et al. [244] presents the AutoDebug framework for automatically sampling and updating several prompts for use in adversarial testing of the language model. Yu et al. argue that evaluating information-retrieval LLMs on popular datasets like Wikipedia provides an over-approximation on the accuracy of these models since they have been overfit on the same data, possibly leading to memorization. Thus, the authors specifically explore adversarial testing under the case that the model predicts a correct response when provided relevant context, but generates an incorrect prediction when the evidence is modified. They apply two different modification approaches: replacing tokens within the context to provide incorrect facts, and adding additional relevant facts to the prompt that may make it difficult to pick out the most important details. The authors collect a new dataset by applying their adversarial generator to Natural Questions [103] and RealtimeQA [94], with additional human filtering to ensure plausible answers are collected.
- (3) The work from Ramakrishna et al. [170] is described in Appendix C.3.2.3.
- (4) The work from Uluoglakci and Temizel [209] is described in Appendix C.3.2.4.

All in all, adversarial prompting is an effective method for identifying robustness of models to unseen inputs, which can be used to develop stronger input filters or fine-tune the model for decreased hallucination tendency.

**4.3.3 Proxy Model.** Certain black-box works rely on an external, proxy model to detect and mitigate hallucinations.

- (1) One such method is used as a baseline within the SelfCheckGPT article [136]. As many language foundation models do not provide access to token probabilities, the authors use an open-source proxy LLM that does provide token probabilities as an estimate of the original output’s probability. They find that using proxy LLMs for probability estimation and hallucination detection successfully is highly variable. The accuracy of detection is dependent on the complexity of the LLM itself, as well as the training data of the proxy LLM (*i.e.*, models trained on independent datasets from the original LLM will have different generation patterns). Refer to the description of Method ID 4.3.1.1 for a discussion on the custom dataset used for evaluation.
- (2) Within this section, we also include works that use an external trained classifier to detect hallucinations. For example, Chen et al. [33] curate a dataset of QA dialogue from LLM generated responses. They apply a composition of metrics to assess quality of responses, including a self-assessment from the LLM comparing the ground-truth and predicted text, human-labeled, and machine metrics (*e.g.*, BERTScore [252], F1 score, BLEU [153], *etc.*). Their hallucination discriminator, RelD, is trained on the dataset in multiple separate phases, each using a different objective: regression, multi-class classification, and finally binary classification. Through experiments, they find that RelD closely aligns with human evaluators’ original predictions.
- (3) The work from Pacchiardi et al. [152] is described in Appendix C.3.3.3.
- (4) The work from Chu et al. [36] is described in Appendix C.3.3.4.

**4.3.4 Grounding Knowledge.** In knowledge grounding tasks, a language model is tasked with identifying evidence from an external knowledge-base that supports claims within a summary. Although seemingly irrelevant to decision-making scenarios, similar methods to ones discussed in this section may be applied in planning tasks to identify observations that are most relevant to predicting the next action, or to generate reasoning behind a specified plan.

- (1) PURR, proposed by Chen et al. [26], is a denoising agent, like LURE (discussed in Section 4.2.1), that is trained in an unsupervised fashion given evidence from online sources, a clean (correct) summary, and a noisy (hallucinated) summary. The model learns to denoise the incorrect summary to the clean statement. During deployment, given a possibly hallucinated claim, a question generation model queries online sources for evidence about the claim, and PURR generates a cleaned version of the original summary with said evidence.
- (2) The work from Li et al. [118] is described in Appendix C.3.4.2.
- (3) The work from Zhang et al. [251] is described in Appendix C.3.4.3.
- (4) Peng et al. [157] aim to add plug-and-play modules to an LLM to make its outputs more accurate, since these large foundation models cannot feasibly be fine-tuned whenever there is new information. Their work formulates the user conversation system as a Markov decision process (MDP) whose state space is an infinite set of dialogue states which encode the information stored in a memory bank, and whose discrete action space includes actions to call a knowledge consolidator to summarize evidence, to call an LLM prompt engine to generate responses, and to send its response to the user if it passes verification with a utility module. The proposed LLM-Augmenter has a memory storing dialogue history, evidence from the consolidator, set of output responses from an LLM, and utility module results. Its policy is trained in multiple phases with REINFORCE [222] starting with bootstrapping from a rule-based policy designed from domain experts, then learning from simulators, and finally, from real users. Peng et al. deploy LLM-Augmenter to two different information retrieval domains: news and customer service. For the news application, the authors retrieve relevant news articles by crawling Reddit news forums for 1.3K articles, following the DSTC7 Track

2 [242] approach. Similarly, to emulate the required knowledge of a customer service chat bot, the authors use the data from DSTC11 Track 5 [98], which holds 14.7K examples of user reviews and frequently answered questions. The authors find that access to ground-truth knowledge drastically improves QA results, and feedback from the utility module and knowledge consolidator help to provide more accurate answers.

- (5) Evaluated in decision-making settings, Introspective Tips [30] provide concise, relevant information to a language planner to learn to solve problems more efficiently. Intuitively, summaries that collect information over all past experiences may be long and contain unnecessary details. In contrast, tips are compact information with high-level guidance that can be learned from one's own experiences, from other demonstrations, and from other tasks in a similar setting. Chen et al. show that providing low-level trajectories is less effective than tips on simulated planning tasks. Additionally, with expert demonstrations, the LLM learns faster with fewer number of failed trials than with just past experience alone. However, one limitation identified in the study is that the LLM underperforms in unseen, low-difficulty missions where it has issues generating general tips for zero-shot testing.
- (6) The work from Lim and Shim [123] is described in Appendix C.3.4.6.
- (7) The work from Schramowski et al. [182] is described in Appendix C.3.4.7.

**4.3.5 Constraint Satisfaction.** There is also additional work in creating black-box algorithms for ensuring decision plans generated by foundation models meet user-defined goal specifications and system constraints, like their grey-box counterpart developed by Wang et al. [215].

- (1) Because these models under test provide their results in text form, it is natural to apply formal method approaches (e.g., satisfiability modulo theory, SMT, solvers) to verify the satisfaction of generated plans. For example, Jha et al. [89] prompt an LLM planner with a problem formulated with first order constraints to predict a set of actions to complete the task. The output plan is input to an SMT solver to check for any infeasibilities in the program, and any counterexamples found are used to iteratively update the prompt and generate new plans. This counterexample approach is much faster than relying on combinatorial search methods that find a plan from scratch. However, the quality of generated plans and the number of iterations before a successful plan is generated are heavily dependent on the LLM generator itself, with similar reasons to the proxy-model used by Manakul et al. [136]. In particular, the authors explore the capability and efficiency of state-of-the-art large language models to solve block-world planning tasks [69]. For every experiment, each LLM is fed the initial random setup of a finite number of blocks in a scene, and the desired goal setup in the form of first-order constraints. Inoperable plans are detected by the Z3 SMT solver [45], who iteratively works with the LLM to approach a feasible solution using counterexamples.
- (2) Another work from Hu et al. [80] develops a RoboEval benchmark to test generated plans on real robots, in a black-box manner. Like Wang et al. [215], the authors introduce their own extension of LTL formulations, known as RTL, which specifies temporal logic at a higher, scenario-specific, level, while abstracting away constraints that are not dependent on available robot skills. RTL and LTL-NL are easier to read and define than classic LTL methods. RoboEval utilizes the provided RTL formulation of a problem, a simulator, and evaluator to systematically check whether the output meets requested goals. Furthermore, to check for robustness of the model to varied instructions, Hu et al. hand-engineer paraphrased sentences within an offline dataset that should ideally result in the same task completion. Primary causes of failures were found to be a result of generated code syntax errors, attempting to execute infeasible actions on the robot, and failing RTL checks.

Like adversarial prompting approaches, testing generated plans on robots in diverse scenarios enable researchers to design more robust systems that hallucinate less frequently at test-time.

## 5 Guidelines on Current Methodologies

In section 4, we present a taxonomy of hallucination detection and mitigation algorithms in various deployment settings. Combining our findings from our extensive review, we now present guidelines for choosing hallucination intervention algorithms and metrics for different environments. As such, we hope these guidelines will enable developers to follow a standardized procedure to define hallucinations for their deployment context, design intervention algorithms, and evaluate efficacy before integrating hallucination-prone models into systems with humans at risk. Even while the field of LLMs evolves rapidly, we argue our guidelines are general enough to continue to assist researchers in the near future. Additional considerations when using deep learning methods are discussed in Appendix E.1.

### 5.1 Process of Choosing and Integrating Intervention Algorithms

We split the design process of hallucination detection and mitigation methods into nine steps, shown in Figure 2.

*Describe the Model Under Test.* The model under test is the specific L(V)LM that has the potential to hallucinate in a deployment environment. Describing the model requires understanding its space of inputs and outputs, and whether designers have access to model weights and/or generation probability distributions. Understanding the model under test is important for narrowing down possible hallucination intervention methods.

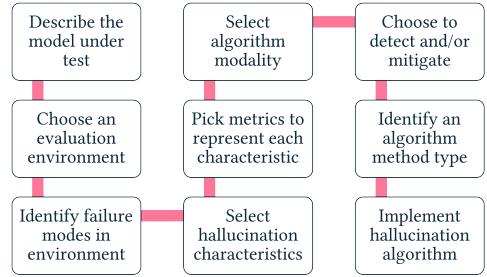


Fig. 2. Design process of hallucination intervention methods.

*Choosing an Evaluation Setting.* Evaluation settings are datasets, simulators, or real-world environments where the model under test and hallucination intervention algorithm are being tested hand in hand. Several possible evaluation settings are described in Appendix D, categorized by application area (e.g., image captioning, conversational QA, control, code generation). Developers should choose an evaluation setting that most closely resembles the intended deployment area to minimize the sim-to-real gap between evaluation and final integration [74, 179, 255]. A too large of a sim-to-real gap results in a poor understanding of the true capabilities of the deployed model. However, designers need to balance the tradeoffs between the cost of developing a new high-fidelity simulator or collecting representative offline data, and using off the shelf low-fidelity environments or datasets that may not truly cover the distribution of the intended deployment area.

*Identifying Failure Modes in Evaluation Setting.* Failure modes are directly related to the characteristics of hallucinations that need to be detected or mitigated. These modes are identified in one of three primary ways. Firstly, designers can choose failure modes based on the hard constraints of the evaluation setting. For example, in decision-making contexts, failure modes may represent a collision space that a robot should not enter, and in QA settings, failure modes could be defined by a set of ground-truth answers that a model generates conflicting results against. Second, stakeholders impacted by model decisions will have personal preferences for how the model under test should act in deployment. As such, the next set of failure modes are defined by acting outside of the range

of desired behaviors. Finally, as discussed in Section 1, it is nearly impossible to account for all possible failure modes at inference time. Thus, intervention algorithm designers should also query the L(V)LM under test with different inputs to identify other undesired and irrelevant generations. Furthermore, many models are released alongside with a *model card* with details of failures found during its design stage [66, 140, 148], which developers can use to identify additional modes to consider. This final sampling procedure will not cover all possible remaining failure modes, but will improve coverage rate. Through this three step process, engineers gain an understanding of what behaviors should be considered as hallucinations in the particular evaluation setting. Although generalized detection/mitigation methods are preferred, identifying these specific subsets of hallucinations will allow engineers to quantitatively evaluate the efficacy of intervention approaches prior to deployment.

*Selecting Hallucination Characteristics.* Given that designers have selected a set of failure modes, or hallucination behaviors, they can now categorize each mode as one of three content characteristics defined in Section 3.1: compliance, desirability, and relevancy. Examples of categorizing failure modes are shown in Table 2. It might be the case that one or more of the categories do not end up with any failure modes. In this case, remaining unmatched characteristics are deemed irrelevant to the development of the hallucination intervention algorithm at this stage. Finally, if stakeholders also care about cases that the L(V)LM is attempting to deceive them through syntactically plausible generations, the plausibility characteristic is also relevant to hallucination detection or mitigation.

*Picking Metrics per Characteristic.* Intervention algorithm designers can now choose metrics to measure the efficacy of model generations and hallucination detection/mitigation methods in meeting the defined requirements of each characteristic. Metrics for each chosen characteristic will differ depending on the deployment context and outputs of the model under test. As such, we provide examples of current metrics in literature categorized by characteristic and application area in Table 4 from our review in Section 4.

*Select Intervention Algorithm Modality.* By this point, designers have performed the prerequisites of describing the available information from the model under test, its limitations, and defining relevant hallucination characteristics and metrics. The remaining steps are actually developing the intervention algorithm. To do so, engineers need to choose which modality the proposed algorithm will fall under (*i.e.*, white-, grey-, or black-box). As described in Section 4, possible modalities depends on the availability of access to necessary information from the model under test. It is also important to consider at this stage the desired flexibility of the intervention algorithm (*i.e.*, the ease of integrating the algorithm to different test models and evaluation settings). For example, black-box methods are easier to deploy to proprietary models where access to weights and token probabilities is limited. However, if the designer only needs to intervene in the hallucination tendency of a particular open-source language model, white-box intervention algorithms can be tuned to the limitations of that specific model.

*Intervention Type.* Next engineers can choose whether the intervention algorithm should detect and/or mitigate hallucinations. Again, this choice depends on the intended deployment area of the model under test and the needs of the stakeholders in the model design process. As seen in Table 3, we find several existing algorithms enable both detection and mitigation of hallucinations. Detection of hallucinations enables reacting to failures, while pure mitigation relies on the efficacy of the intervention algorithm alone to proactively filter hallucinations prior to final generation. As such, we recommend safety-critical scenarios be deployed with detection and mitigation algorithms to simultaneously reduce chances of failures and inform impacted parties of potential hallucinations to increase transparency.

**Table 4. Common datasets/simulators and metrics across settings and hallucination characteristics.** Each metric is accompanied by references that use the metric as the specified characteristic. The same metric could be applied as different characteristics.

Setting	Common Datasets/Simulators	Characteristic			
		Compliance	Desirability	Relevancy	Plausibility
QA	– Accuracy <sup>†</sup> [7, 26, 33, 49, 56, 58, 102, 118, 136, 144, 152, 157, 229, 241, 244, 251],				
	– Contradictions [49, 56, 58, 144, 152, 241],	– Calibration Error [124, 229],			
	– Concept Probabilities [136, 165, 210],	– Succinctness [157, 194],			
	– TriviaQA [92],	– Over-Conservativeness [234],	– Perplexity [144, 194],	– ROUGE [157, 165, 170],	
	– GSM8K [38],	– # of Clarifying Questions [251],	– Cross Encoder [26]	– Semantic Preciseness [33, 194]	
	– HotpotQA [237],	– Prudence [234],			
	– MMLU [76],	– Preservation [26],			
	– Natural Questions [103]	– TOXIGEN Classifier [138]			
	[238, 244],				
	– Coverage [102, 251],				
IC	– METEOR [157, 170],				
	– SummaC [58],				
	– FactScore [49],				
	– Cosine Similarity [241],				
	– Sentence-Bert [244],				
	– AlignScore [170]				
	[259],				
	– LVLM-Based Scoring [82, 119, 259],				
	– CHAIR [82, 119, 259],				
	– POPE <sup>†</sup> [82, 119, 259],				
IG	– MSCOCO [126],	– Co-Occurrence [259],	– CIDER Human Alignment [259],	– CHAIR [82, 119, 259],	– Perplexity [82],
	– GQA [85],	– Concept Probabilities [259],	– Detailedness [82]	– POPE <sup>†</sup> [82, 119, 259]	– ROUGE [259],
	– Visual Genome [100],	– BLEU [259],			– SPICE [259]
	– A-OKVQA [185]	– BertScore [259],			
		– CLIP Score [259],			
		– METEOR [259]			
	[138, 182],				
	– NudeNet Classifier [138, 182],				
	– Q16 Classifier [138, 182],				
	– Diversity [138],				
P	– MSCOCO [126],	– Contradictions [36, 123],	– Monetary Cost [36],	– LVLM-Based Scoring [36],	– Factual Fabrications [123]
	– LSUN [243],	– Accuracy <sup>†</sup> [36],	– Bias [182],		
	– LAION-400M [184],	– LVLM-Based Scoring [36],	– Inappropriate Probability [182],		
	– I2P [182]	– CLIP Score [182]	– Expected		
			Inappropriateness [182],		
			– COCO FID-30k Fidelity [182]		
	[30, 73, 80, 89, 122, 154, 174, 215],				
	– Generalizability [30, 73, 122, 154],				
	– Clarification/Unsure Rate [122, 154, 174, 215],				
	– Feasibility [73, 89, 122, 154, 174, 215],				
†	– RoboEval [80],	– Action Probabilities [73, 122, 174],	– # of Attempts [30, 80, 89],	– Action Probabilities [73, 122, 174]	– Non-Compliance Rate [122],
	– SaGC [154],	– Plan/Action Accuracy <sup>†</sup> [80, 154, 215],	– Calibration Error [122, 174],		– Preciseness [122]
	– Ravens [246],	– Action Variance [182],	– Coverage [122, 174],		
	– BabyAI [34],	– Contradictions [89]	– Estimated Payoff [30, 73],		
	– TableSim [174]		– Action Set Size [122, 174],		
			– Plan Readability [215],		
			– Plan Length [73],		
			– Unsafe Action Rate [122],		
			– Monetary Cost [122],		
			– Inference Speed [215]		

<sup>†</sup>Encompasses machine metrics like (balanced) accuracy, precision, recall, F1, AUC, exact match [28], and pass@1 [31].

**Table 5. Benefits and limitations of each intervention algorithm type.**

Method Type	Pros	Cons
Hidden States	<ul style="list-style-type: none"> <li>– tuned for specific model</li> <li>– hidden states hold useful embeddings</li> <li>– no need to re-embed text output</li> </ul>	<ul style="list-style-type: none"> <li>– reduced model transfer flexibility</li> <li>– not applicable for proprietary models</li> </ul>
Attention Weights	<ul style="list-style-type: none"> <li>– tuned for specific model</li> <li>– attention weights hold useful info</li> </ul>	<ul style="list-style-type: none"> <li>– reduced model transfer flexibility</li> <li>– not applicable for proprietary models</li> </ul>
Honesty Alignment	<ul style="list-style-type: none"> <li>– directly fine-tunes model under test</li> <li>– empirically generalizes to new data</li> </ul>	<ul style="list-style-type: none"> <li>– tuned model still susceptible</li> <li>– test efficacy impacted by data quality</li> </ul>
Concept Probabilities	<ul style="list-style-type: none"> <li>– generally model agnostic</li> <li>– intuitive</li> <li>– tried in broadest set of deployments</li> </ul>	<ul style="list-style-type: none"> <li>– requires access to token probabilities</li> <li>– correlation may not necessarily hold</li> </ul>
Conformal Prediction	<ul style="list-style-type: none"> <li>– theoretical guarantees</li> <li>– applicable to multi-step planning</li> <li>– provides model uncertainty metric</li> </ul>	<ul style="list-style-type: none"> <li>– requires access to token probabilities</li> <li>– requires collecting calibration dataset</li> <li>– relies on human intervention</li> </ul>
Analyzing Samples	<ul style="list-style-type: none"> <li>– applicable to proprietary models</li> <li>– intuitive</li> <li>– removing conflicts reduces failures</li> </ul>	<ul style="list-style-type: none"> <li>– efficiency impacted by # of samples</li> <li>– affected by compliance metric choice</li> </ul>
Adversarial Prompting	<ul style="list-style-type: none"> <li>– applicable to proprietary models</li> <li>– reveals undesired behaviors</li> <li>– can lead to better filters</li> <li>– red-teaming often occurs pre-release</li> </ul>	<ul style="list-style-type: none"> <li>– requires covering large input space</li> <li>– generally does not mitigate</li> <li>– requires hallucination classifier</li> </ul>
Proxy Model	<ul style="list-style-type: none"> <li>– applicable to proprietary models</li> <li>– simple classifier could work well</li> </ul>	<ul style="list-style-type: none"> <li>– proxy has mismatched distribution</li> <li>– dependent on proxy complexity</li> <li>– LLM proxy could hallucinate</li> </ul>
Grounding Knowledge	<ul style="list-style-type: none"> <li>– applicable to proprietary models</li> <li>– provides evidence for responses</li> <li>– aligns model knowledge with users'</li> <li>– can help to learn policies faster</li> </ul>	<ul style="list-style-type: none"> <li>– requires ground-truth database</li> <li>– could reference stale knowledge</li> <li>– still fails in low-data regimes</li> </ul>
Constraint Satisfaction	<ul style="list-style-type: none"> <li>– applicable to proprietary models</li> <li>– theoretical guarantees</li> <li>– SMT solvers find failure cases quickly</li> </ul>	<ul style="list-style-type: none"> <li>– requires precise constraint definitions</li> <li>– specification may be hard to parse</li> </ul>

*Identifying an Intervention Sub-Type.* Now that the deployment setting, modality, and intervention type have been identified, engineers can choose a method type from Table 3 that falls within those constraints. This step will require some experimentation across method types and specific algorithms to identify the most effective intervention approach using the previously chosen metrics. We list pros and cons of each method type in Table 5 to assist researchers with narrowing down the scope of their algorithm search.

*Implementing the Hallucination Intervention Algorithm.* Finally, designers can implement and integrate a chosen algorithm into the specific deployment setting and perform additional tests to measure its efficacy in detecting/mitigating hallucinations from the model under test.

## 6 Future Directions

Here, we discuss some possible future directions in hallucination detection and mitigation techniques for foundation models to improve deployments to decision-making tasks.

*Evaluating Methods on Decision-Making Tasks.* Most hallucination detection approaches are currently tested in offline QA settings for information retrieval or knowledge alignment, as seen in Table 3. As foundation models are increasingly used for more complex tasks, researchers should make an effort to adapt and evaluate earlier detection/mitigation approaches that were applied to QA problems. Although dissimilar in practice from QA settings, planning and control problems may be formulated such that earlier mitigation methods can be evaluated on decision-making tasks. For example, as discussed in Section 2.1, Chen et al. [29] treat the autonomous driving task as a QA

problem, which could be naturally extended to test other QA hallucination detection methods in the same setting. This evaluation may lead to greater understanding of the general limitations of these models, as we draw parallels across diverse deployments.

*Development of More Black-box Approaches.* White- and grey-box detection methods may not generally be applicable in situations where the internal state or token probabilities are unavailable from the language model. Thus, we predict black-box approaches will take precedence in the near future, as state-of-the-art LVLMs like GPT-4V already prohibit access to probability outputs. However, current black-box methods are limited with simplistic sampling techniques to gauge uncertainty, and proxy models may not be representative of the true state of the model under test. Works like FLIRT showcase the promise of black-box adversarial prompting approaches in generating undesirable results [138]. We argue developing more aggressive black-box adversarial generative models, which explicitly optimize for producing inputs that may perturb the system outputs, is key to identifying the limits of a foundation model’s knowledge.

*Pushing Models’ Generalization Capabilities.* Currently, foundation models are primarily deployed to decision-making tasks that likely have some relation to its training set. For example, although complex, tasks like multi-agent communication, autonomous driving, and code generation will be present in training datasets. On the other hand, dynamic environments like robot crowd navigation require identifying nuances in pedestrian behaviors which the model may not have explicitly seen during training. Thus, when models are deployed in decision-making contexts and encounter a previously unseen scenario, or they are utilized in a completely different setting from their training data, it is necessary to consider their generalization capabilities. As discussed in Section 2.1, existing examples of foundation models applied in autonomous driving and robotics utilize external tools to retrieve sensor data or memories before planning. Mialon et al. [141] refer to such models that extract useful information from databases as augmented language models. As such, the authors argue that the combined efforts of using external tools and internal reasoning is critical to the generalizability of language models to broader tasks — also explored by Chen et al. [30], Park et al. [155], and Wang et al. [213]. From another perspective, Tong et al. [205] approach the development of generalized multi-modal large language models from a vision-centric focus. In particular, the authors find that training LVLMs with heavy consideration on the design of vision encoders and the respective connector between the vision and language models drastically improves the capabilities of architectures deployed in vision tasks (e.g., image captioning, QA, depth ordering). This line of thinking can bolster the performance of LVLMs deployed in autonomous driving decision-making, where current methods have failed [221]. Before integrating models into real-world applications, we argue that designers should thoroughly explore their generalization limitations to find directions for future growth and to maximize transparency of model capabilities.

*Testing Multi-modal Models.* With the explosion of LVLMs, which allow for explicit grounding of natural language and vision modalities, further exploration should be performed in evaluating their effectiveness in decision-making systems. Wen et al. [221] take a step in the right direction towards testing black-box LVLMs in offline driving scenarios, but there is still work to be done in deploying these models in online settings. This direction can shed light on the long-standing debate of whether modular or end-to-end systems should be preferred in a particular deployment setting. In fact, while our work has focused on LVLMs, there exist other families of multi-modal foundation models for the audio [151] and 3D generation [32] spaces, which similarly hallucinate [178, 214] and should be evaluated before deployment.

*Summary.* We provide a glimpse into the progress of research for evaluating hallucinations of foundation models for decision-making problems. First, we identify existing use cases of foundation models in decision-making applications (e.g., autonomous driving, robotics) and find several works make note of undesired hallucinated generations in practice. By referencing works that encounter hallucinations across diverse domains, we provide a flexible definition for hallucinations that researchers can leverage, regardless of their deployment scenario. Then, we give a taxonomy of existing hallucination detection and mitigation approaches for decision-making, question-answering, etc., alongside a list of commonly used metrics, datasets, and simulators for evaluation. We find that existing methods range in varying assumptions of inputs and evaluation settings, and believe there is room for growth in general, black-box hallucination detection algorithms for foundation models. Finally, we present generalized guidelines to assist engineers with selecting hallucination intervention algorithms across varied deployment contexts, and suggest future research directions.

## A Foundation Models Making Decisions

### A.1 Robotics

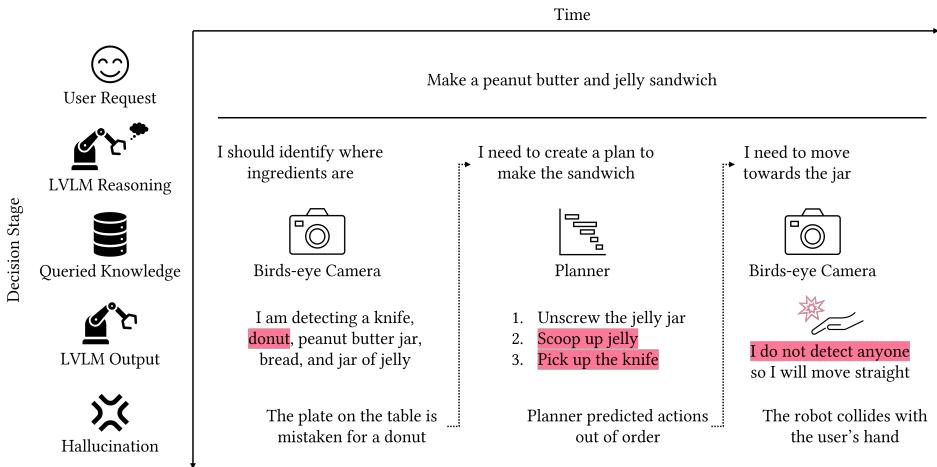


Fig. 3. **Example deployment of an LVM foundation model in a robotics setting.** Hallucinations are highlighted pink. Here, a robot tasked with assembling a sandwich initially identifies an object incorrectly. Then, the model comes up with an infeasible plan. Finally, when attempting to perform one of the actions, the robot collides with the human as it did not perceive any danger.

### A.2 Other Areas

There are also other works that apply foundation models for decision-making outside of the robotics and autonomous vehicle domains. For example, ReAct from Yao et al. [239] identifies that a key limitation of chain-of-thought reasoning [217] is that the model does not update its context or action based on observations from an environment. As such, chain-of-thought reasoning relies purely on the internal reasoning of the foundation model itself to predict actions to take, missing a crucial step in grounding its actions with their effects on the environment. Given a prompt, ReAct iterates between an internal reasoning step and acting in the environment to build up context relevant to the task. Yao et al. showcase the promise of the method in a QA setting where the LLM

can take actions to query information from an external knowledge base, as well as an interactive text-based game, ALFWORLD [190]. Chen et al. [30] admit that ReAct is a powerful tool for dynamic reasoning and grounding, but is limited by the fact that the updated context from the Act step is only helpful for the particular task the model is currently deployed for. They propose Introspective Tips to allow an LLM to reason about its past successes and failures in a world to generate general tips that will be helpful across diverse instruction-following tasks. Specifically, tips are generated from the past experience of the model from a similar set of tasks, from expert demonstrations, and from several games that differ from the target task. By summarizing these experiences into more concise tips, Chen et al. show that Introspective Tips outperform other methods in ALFWORLD with both few- and zero-shot contexts.

Park et al. [155] and Wang et al. [213] apply foundation models in more complex environments to push models to their limits to simulate realistic human behaviors and test lifelong learning. Park et al. propose generative agents that produce believable, human-like interactions and decisions within a small town sandbox environment. They develop a module for individual agents in the simulation to store and retrieve memories, reflect about past and current experiences, and interact with other agents. Their generative agents use similar methods to ReAct and Introspective Tips to act based on a memory of experiences, but also interact and build relationships with other agents through dialogue. The authors show that the agents are able to effectively spread information, recall what has been said to others and stay consistent in future dialogue interactions, and coordinate events together. Sometimes, however, agents are found to hallucinate and *embellish* their responses with irrelevant details that may be attributed to the training dataset or outside, real-world knowledge. Voyager, from Wang et al., deploys GPT-4 to the MineDojo environment [60] to test its in-context lifelong learning capabilities. The architecture prompts GPT-4 to generate next high-level tasks to complete, given the agent’s current state and results of past tasks — a form of automatic curriculum generation. Voyager then identifies what intermediate general skills would be required to complete the task, and the LLM is used to fill in a skill library with helpful low-level skills in the form of programs that call functions that are available to the simulator. GPT-4 is prompted to generate skills that are generalizable to multiple tasks, so that the skill generation step does not have to be called for every task if the skill is already stored in the library. Wang et al. show that Voyager continuously learns to explore the diverse tech tree available within MineDojo while building and leveraging skills. Even so, they find that the LLM hallucinates when generating tasks to tackle and when writing the code to execute for a particular skill, discussed further in Section 3.2.

Kwon et al. [104] explore the use of LLMs to act as a proxy for a hand-tuned reward function in RL tasks. This application is particularly motivated by decision-making tasks that are difficult to specify with a reward function, but can be explained textually with preferences of how a policy should generally act. Specifically, the LLM evaluator first undergoes in-context learning with examples of how it should decide the reward in several cases of the task that the agent will be deployed to. Then, during RL training, the LLM is provided a prompt with the trajectory of the agent within the episode, the resulting state from the simulator, and the original task objective from the user, and is asked to generate a binary reward for the agent (1 if success, 0 else). The binary reward is added to the experience replay, and the agent can be updated using any RL algorithm. Kwon et al. find that a baseline in their work that predicts rewards with no in-context learning especially hallucinates with incoherent reasoning. There has also been work by Suri et al. [199] on evaluating whether large language models elicit heuristics and biases when making decisions, like humans. Specifically, while it is commonly believed in literature that biases like anchoring, representativeness, availability, framing, and endowment are brought about by cognitive processes, Suri et al. pose that if foundation models generate similar responses, these effects may be partly caused by language, since language models inherently have no cognitive capabilities.

Through four user studies comparing the responses of humans and ChatGPT to the same questions intending to bring about the decision biases, the authors find that both the model and participants showcased similar effects of partiality. As such, the authors posit that these results imply that decision heuristics in people may actually be influenced by linguistic syntax. We further argue in Appendix B.2 that language models deployed in decision-making contexts where they may impact humans' lives should aim to minimize these biases for fairness.

## B Hallucinations

### B.1 Examples

*Image and Video Generation.* Likewise, image and video generation models are not safe from hallucinations. Recent generative models like DALL-E [15, 171], Stable Diffusion [176], and SORA [149] have shown great promise in producing high quality frames given text input, but they can also hallucinate noncompliant, undesirable, and irrelevant features. The DALL-E system cards [143, 148] provide plenty of examples of such characteristics with biased, stereotypical generations, deepfakes of public figures, and violent and racy imagery. Betker et al. [15], who train DALL-E 3 to generate images given synthetic captions from a captioning model, find that poor, hallucinated training captions result in noncompliant image generations at test time that ignore important caption details. Furthermore, generations may also include nonfactual contents that misguide users [123]. Even with the development of safety filters that attempt to detect hallucinations before displaying the generation to the user, Rando et al. [172] find that simple prompt engineering and red-teaming approaches can still lead to vulgar content, bypassing the filter. As such, recent efforts have shifted from solely filtering inappropriate content at test time, and instead providing guidance to the model to assist in generating safer features [182]. However, instances of stereotypes are still prevalent in generations. In a positive light, Huang et al. [83] propose exploiting the hallucination tendency of image generation models to curate a benchmark to evaluate image captioning methods. While the results of image generation models should minimize hallucinations in static frames, video generation models also need to produce temporally consistent and plausible clips across a sequence of frames. Sora Detector, proposed by Chu et al. [36], attempts to detect static hallucinations (e.g., color distortion, deformations, unrealistic depth of field) and dynamic hallucinations (e.g., unnatural overlapping of objects, implausible motions, temporally illogical generations) within generated videos conditioned on text.

*3D Modeling and Generation.* Another field that is quickly gaining the attention of researchers is 3D generation, where foundation models are generating high fidelity representations of objects or scenes conditioned on images or text [9, 114, 129, 227]. These foundation models span a wide range of downstream applications including human avatar creation [230], medical tomography understanding [16], 3D object generation [32, 70], environment modeling [235], and more. Although billions of labeled pairs of images and text exist across many datasets, as discussed in Appendix D.2, 3D modality labels have only recently begun to approach a similar scale [232], limiting the number of training pairs for 3D foundation models. As such, early 3D generative foundation models conditioned on text relied on embeddings from text-image similarity models like CLIP [167] to serve as a guide for generating accurate scenes [87, 180], or used pre-trained text-to-image diffusion models to provide feedback for training 3D neural radiance field models [160]. Instead of optimizing individual low-level object representations directly, some works assume access to a collection of 3D assets, which a standard LLM like GPT-4 can query to complete the floor plan of a 3D scene, bypassing the need for low-level text-3D datasets [235]. Yang et al. [232] are some of the latest researchers to tackle the data scarcity problem of text-labeled 3D scenes by introducing their own procedure for collecting text-scene labels across 40K rooms. When evaluating models with

their new 3D-POPE metric, the authors find that several 3D LLMs hallucinate the presence of non-existent objects in the scene, similarly to other LVLMs tested on images. Wang et al. [214] find that 3D generative foundation models hallucinate spatial inconsistencies when rendering a generated scene from different perspectives. Due to the scarcity of 3D datasets, the authors propose Hallo3D: a three-part hallucination detection and mitigation approach leveraging LVLMs as advisors. Specifically, Hallo3D samples multiple rendered views of a 3D object with a pre-trained diffusion model conditioned on text. Each rendering is passed through an LVLM which identifies abnormalities, and the predicted statements are fed into the diffusion model as negative contexts to update the original renderings through a *prompt-enhanced reconsistency* step. The authors use a cross-attention mechanism to improve the consistency of frames across different view points.

## B.2 Broader Impacts

Thus far, we have primarily formulated a unified definition for hallucinations and provided examples of where they have come up in the research community. However, as this new field advances rapidly every day, numerous industries have begun utilizing the technology in their respective applications. As such, it is even more imperative that designers develop robust hallucination detection and mitigation methods before deploying models into areas with real humans at risk. Three critical industries where LVLMs have shown great promise are the medical, legal, and finance sectors [37, 106, 120]. For example, Li et al. [120] and Zhao et al. [254] explain that LLMs are currently used in finance for portfolio management, fraud detection, credit scoring, text summarization for forecasting, and customer-facing chatbots. In fact, Bloomberg has already designed its own LLM, BloombergGPT, trained on a custom curated dataset for investor sentiment analysis given news transcripts, numerical reasoning QA, and named entity recognition [226]. Similarly, LVLMs are being applied to medical tasks for patient education through QA, assisting physicians when writing reports and examining test results, and used in academia for taking medical exams — even going so far as accomplishing passing results [146]. Likewise, Lai et al. [106] showcase recent examples of use cases of LLMs in law, including summarizing legal documents, generating drafts of legal documents, acting as a legal consult, and making decisions given case facts.

For all of their promising results, foundation models have even greater impacts in these critical applications. Specifically, designers need to consider problems of data privacy, training with out-of-date data, potential inconsistencies between generations and references, model bias, the ethics of using AI for a particular problem setting, and transparency to stakeholders and impacted parties in the decision-making pipeline. In the finance sector, Kang and Liu [93] identify common failure modes of various open-source and proprietary language models in tasks for recognizing financial abbreviations and stock symbols, providing explanations for financial terms, and querying a stock price without access to an external database. Each model is shown to have defects at inference time without using retrieval augmented generation (RAG) [113] methods (similar methods are discussed in Section 4.3.4), especially when the training data is out of date. As such, companies like BlackRock, Inc. are directly applying RAG to existing LLMs for financial QA [181]. In the context of legal applications, Dahl et al. [43] provide a taxonomy of the complexities of different problems a language model could be tasked with, each with increasing risk of generating undesired hallucinations. Notably, the frequency of hallucinations are found to increase with task complexity and varies with the specified court, jurisdiction, case prominence, and year. In fact, LLM hallucinations have already made their mark in a real court case, where a New York attorney utilized ChatGPT to generate a brief, which in turn referenced non-existent cases. He states the model told him its generations were accurate, underscoring the importance of developing transparent models [218]. Ali et al. [4] evaluate the efficacy of foundation models for automatically generating billing codes given medical

procedure descriptions. As the tested models currently perform miserably on the given task, the authors suggest further research before deploying the models in an automated medical context.

In an even more generalized medical setting, multiple LLMs are queried to answer novel questions, where even RAG results in generations that are either inconsistent with sources, or the cited text cannot be found in the source [225]. In fact, the AI search engine company Perplexity AI, which uses RAG to generate summaries, came under fire during the 2024 United States presidential election for releasing an election hub that generated hallucinated results [39]. One human tester found that, while the company's product centralized information into one source, the model's generations wrote in varied tones for different candidates, resulting in biased generations. Similarly, LLMs have been found to generate fake citations to support claims queried by users. For example, an education official from the state of Alaska used a generative model to collect citations for a proposed policy [197]. The burden was then left to the policymaker to replace citations to nonexistent sources, leading journalists to question the lack of policy on using generative AI to write proposals which impact the public. The promise of LLMs have also led many educators to directly rely on these models to produce lesson plans, leading to hallucinations when left unchecked [12]. One audio transcription tool leveraged in medical domains has been found to add irrelevant details to generations, and erase the ground-truth audio for security [200]. In an era where foundation models are being deployed in new, critical areas, we argue ground-truth data should be kept for later model evaluation and redundancies. Overall, it is of the utmost importance to ensure LVLM generations are free of hallucinations in critical decision-making applications. While these models are still far from reaching this milestone, it is even more important to be transparent about model capabilities to users to minimize misalignment of expectations.

## C Detection and Mitigation Strategies

### C.1 White-box Methods

#### C.1.1 Hidden States.

(3) LUNA, introduced by Song et al. [194], is a general framework that measures the trustworthiness of an LLM output containing four stages of evaluation: model construction, semantic binding, quality metrics, and practical application. The abstract model construction phase attempts to profile the LLM using its hidden states with either a discrete time Markov chain (DTMC) or a hidden Markov model (HMM) architecture. For example, when fitting a DTMC model, the authors encode the hidden states of the language model into a lower dimensional space, cluster them into abstract discrete states, and learn a transition function between said states. Semantic binding is used alongside quality metrics to identify the states and transitions that are trustworthy, and which ones are undesired. Finally, at inference time, as the model generates output tokens to a given prompt, the intermediate network layer embeddings are iteratively passed through the profiling model to identify when undesired transitions occur. The authors evaluate their framework's capability of detecting hallucinations within QA datasets.

#### C.1.3 Honesty Alignment.

(2) Yang et al. [234] take the method one step further by also training the model to refuse to answer questions with high uncertainty.

### C.2 Grey-box Methods

#### C.2.2 Conformal Prediction.

- (2) Kumar et al. [102] similarly apply conformal prediction to LLMs, but for answering multiple choice questions. Specifically, the method first collects a calibration dataset of prompts and the normalized token probabilities of the correct token (*i.e.*, A, B, C, or D) being chosen from the model. Then, during deployment, given a user-defined error rate and a prompt, their algorithm chooses the multiple choice answers with token probabilities that fall within the calibrated score on the held-out dataset.

### C.3 Black-box Methods

#### C.3.1 Analyzing Samples from Model.

- (2) A concurrent work from Elaraby et al. [58] rather computes the *entailment* among responses at the sentence-level. Consistency metrics check whether responses contradict one another while entailment metrics identify if the responses imply one another. The nuanced difference between SelfCheck-NLI and their method, HaloCheck, is that Elaraby et al. use the Summac [105] entailment estimation method, placing equal weightage among all sentences, in all responses, to compute a more balanced prediction score. The authors evaluate HaloCheck and hallucination mitigation techniques on a domain-specific, custom-curated dataset, with facts about the US National Basketball Association (NBA). Specifically, Elaraby et al. first prompt GPT-4 for questions on topics within the NBA domain and manually filter out low-quality generations, resulting in 151 questions. An LLM is then queried for 5 responses to each question, and each response is manually annotated for consistency and accuracy of generations. HaloCheck is also shown to be more efficient at predicting scores. Other commonly used metrics within the language community for similarity estimation are discussed in Appendix D.1.1.
- (7) In the problem space of image captioning, Li et al. [119] estimate the accuracy of a (possibly hallucinating) LVLM when describing an image with a text caption. Early metrics, like CHAIR [175], fail to provide stable estimates of accuracy when different captions with similar semantic grounding lead to varying scores. To tackle this stability problem, the authors propose POPE, which curates binary questions about whether an object exists within an image scene. Questions to which the foundation model provides conflicting responses describe objects that the model may be hallucinating. These metrics are further described in Appendix D.1.2.
- (8) Yet another recent work by Xiong et al. [229] asks LLMs in a zero-shot manner to verbally include their uncertainty in the generated output. This desired behavior is elicited through additional prompt engineering (*i.e.*, add a phrase to the prompt like “Please provide your confidence level as a percentage”). Unlike Lin et al. [124], the authors do not further fine-tune the model to a calibration dataset of uncertainties. To combat over-confidence in output scores, Xiong et al. utilize chain-of-thought reasoning, predict the confidence score of each sub-claim, and combine them over the whole response to compute the final belief. The authors find that a hybrid approach, merging verbalized uncertainty with self-contradiction detection, outperforms the individual components alone on expected calibration error, comparing predicted confidence and actual model accuracy.
- (9) Yehuda et al. [241] present InterrogateLLM, a sampling-based approach that attempts to reconstruct an original query given a possibly hallucinated LLM response, and measures the similarity of the two queries. Inspired by human studies [19], the authors specifically hypothesize that responses that generate queries that differ greatly from the original prompt point to possible hallucinations. However, they also confess that, like human studies, there is the possibility of false positive hallucination detections due to the stochastic nature of

language models. InterrogateLLM follows a simple procedure: form a prompt with few-shot examples containing queries and answers followed lastly by the actual query, generate a (hallucinated) response from a language model under test, reverse the original prompt examples to provide answers and their queries with the generated response appended last, sample generated queries from any language model (not necessarily the one under test), and measure the cosine similarity of vector embeddings between the original query and generated queries. Using public datasets covering a range of trivia knowledge [11, 59, 260], the authors curate a dataset of QA pairs. Most importantly, the authors find that their backward few-shot method is critical to decreasing the false-negative detection rate of hallucinations seen in SelfCheckGPT [136], which only compares generated responses from the original forward pass. Furthermore, using an ensemble of models for the backward generation and more iterations of query sample generation lead to higher detection accuracy, at the cost of efficiency. We expect this backward-query generation approach to hallucination detection can also be applied to planning tasks, where hallucinated plans will reverse-generate task descriptions that do not match well with the original goal.

### C.3.2 Adversarial Prompting.

- (3) Motivated by similar findings that language models are overfitting to QA datasets, Ramakrishna et al. [170] present a novel framework for generating invalid questions to test the frequency of hallucinations output by new LLMs. The authors collect an augmented version of the DBpedia dataset [110] replaced with invalid questions to evaluate the hallucination rate of language models, and note that any deployment dataset could have been used instead. Specifically, Ramakrishna et al. manually create a list of 24 question templates with tags for subjects and objects that can be filled in. A set of 100 invalid questions using the templates is generated by sampling subjects and objects from disjoint facts in the original dataset, and ensuring the questions do not have valid answers in DBpedia. Additional invalid questions are produced by replacing dates within questions from TriviaQA [92] with ones that do not exist. By passing each test question through various open-source and proprietary LLMs, the authors manually validate that each model has a tendency to hallucinate. Unfortunately, Ramakrishna et al. find that using automated evaluation metrics (discussed in Appendix D.1) do not correlate well with human annotations on their generated dataset, leaving room for future growth in aligned automatic evaluation metrics.
- (4) A more recent adversarial method from Uluoglakci and Temizel [209] attempts to automatically generate hypothetical, invalid questions that language models should reject answering. For example, the hypothetical question, “What are the differences between Platypus LLM and Wolf LLM?” should be rejected since Wolf LLM is nonexistent, even though Platypus [108] is a real family of language models. Intuitively, if a model provides a plausible answer to the invalid question, the authors claim that either the fabricated term is not be in the model’s training set or the model has a higher tendency to hallucinate overall. As such, the authors propose a new framework to generate hypothetical questions, with which they create the HypoTermQA dataset. In particular, Uluoglakci and Temizel first query GPT-3.5 for 20 popular topics and then 50 hypothetical terms per topic, resulting in 790 filtered fake phrases. To generate a diverse set of outputs, the model is set with a high temperature parameter. The authors argue that questions produced with only hypothetical terms will be easier to distinguish by language models, and thus, generate a set of similar terms with real meanings, in an attempt to deceive models. Similar, valid phrases were generated using three distinct approaches: querying GPT-3.5 directly, retrieving titles from Wikipedia with similar vector embeddings to

the hypothetical terms, and taking the titles of Wikipedia passages whose definition embeddings are similar to those of the hypothetical terms. Finally, GPT-3.5 is instructed to generate plausible questions with filtered hypothetical and valid terms to produce a total of 19.5K hypothetical and valid questions. Rather than relying on human annotation for identifying hallucinated responses, Uluoglakci and Temizel additionally propose HypoTermQA Score — the ratio of valid answers to the total number of hypothetical questions, automatically labeled by a proxy evaluator LLM agent (like methods in Section 4.3.3). In evaluation, the authors find that both open-source and proprietary models have over 90% frequency of generating invalid responses. However, this frequency differed depending on the chosen evalautor model due to biases. Thus, proxy evaluator agents are also tested by comparing against manual annotation. Overall, the proposed dataset generation and automatic hallucination detection method show great promise in evaluating the factual hallucination tendency of language models. But, additional work should be done to handle biases in evaluator models, evaluate other characteristics of hallucinations, and consider other deployments outside QA.

### C.3.3 Proxy Model.

- (3) Similarly, Pacchiardi et al. [152] develop a black-box lie detector for LLMs. In their case, the authors hypothesize that models that output a lie will produce different behaviors in future responses, like Azaria and Mitchell [7]. As such, at inference time, Pacchiardi et al. prompt the LLM with several binary questions (that may be completely unrelated to the original response) and collect yes/no answers. All the responses are concatenated into a single embedding that is input to the logistic regression model to predict the likelihood that the response was untruthful. To evaluate their lie detector, the authors assemble over 20K questions from existing data sources on topics including general trivia [139, 211, 219], basic arithmetic [156], common sense reasoning [202], text translation [204], and self-awareness [158]. They additionally synthetically generate questions with unknowable answers as a control split (e.g., “What day is it?”). The authors find that the simple detector is mostly task- and model-agnostic once trained on a single dataset.
- (4) Chu et al. [36] primarily rely on the LVLM GPT-4 to detect hallucinations within AI-generated videos. Video hallucinations are classified as static hallucinations, which occur in individual frames, or dynamic hallucinations, which occur across multiple frames. Here, the authors employ the generalized capabilities of GPT-4 to detect objects within keyframes of a video, summarize the video from the keyframes (which might be inconsistent with the original video generation prompt due to hallucinations), synthesize a temporal knowledge graph representing the changing relations among detected objects, detect inconsistencies along the knowledge graph, aggregate a hallucination score, and describe the detected hallucinations. Sora Detector is evaluated on a custom dataset collected by the authors, T2VHaluBench (described in Appendix D.2.4), and outperforms video hallucination detection ablation approaches that ignore knowledge graph construction, showcasing the usefulness of the representation. As such the structured approach shows promise in identifying undesired effects in generated videos, but the authors do not provide discussion on the possible hallucinations generated by GPT-4 when detecting and describing failures.

### C.3.4 Grounding Knowledge.

- (2) Some knowledge grounding approaches prompt LLMs to generate code to directly query information from databases. Li et al. [118] are motivated by the limitations of existing knowledge-based hallucination mitigation methods; namely that (1) they utilize a fixed knowledge source for all questions, (2) generating retrieval questions with LLMs that interface with a database

is not effective because they may not be trained on the particular programming language of the database, and (3) there is no correction capability that handles error propagation between knowledge modules. Consequently, the authors propose augmenting LLMs with heterogeneous knowledge sources to assist with summary generation. Specifically, in the event that the model is found to be uncertain about its generated statement through self-contradiction, their framework, chain-of-knowledge (CoK), chooses subsets of knowledge-bases that may be helpful for answering the original question. Assuming each database has its own query generator, CoK queries for evidence, and corrects rationales between different sources iteratively. Compared to chain-of-thought reasoning, CoK consistently produces more accurate answers with its iterative corrections.

- (3) Another source of potential conflict that leads to hallucinations, is misalignment between a model's capabilities and the user's beliefs about what it can do. Zhang et al. [251] tackle this knowledge alignment problem and categorize alignment failures into four types:

- Semantic – an ambiguous term maps to multiple items in a database
- Contextual – the user failing to explicitly provide constraints
- Structural – user provides constraints that are not feasible in the database
- Logical – complex questions that require multiple queries

Their proposed MixAlign framework interacts with the user to get clarification when the LLM is uncertain about its mapping from the user query to the database. With the original query, knowledge-base evidence, and user clarifications, the LLM formats its final answer to the user.

- (6) Lim and Shim [123] tackle the issue of image generation models producing results that are factually noncompliant with the original prompt, by utilizing an external database to address inconsistencies. In particular, the authors confront three forms of image hallucinations: factual inconsistencies, outdated knowledge, and factual fabrications (unlikely generations). Given a prompt and an original (hallucinated) generation from a generator, the authors manually collect a ground-truth image describing the prompt using Google. An LVLM is employed to identify the differences between the retrieved and generated images, resulting in an instruction for how to edit the original generation to become factual. Finally, the original generation and editing instruction is passed through the generator to produce an updated result. Qualitatively, the pipeline produces more factually compliant and plausible generations. However, a human needs to manually choose the ground-truth image, the prompt correction generation model can hallucinate improper instructions, and the authors do not compare their method with other image hallucination mitigation methods.

- (7) Schramowski et al. [182] present Safe Latent Diffusion (SLD), a method to provide additional guidance to image generation diffusion models at inference time to reduce the tendency of inappropriate generations. Intuitively, diffusion models iteratively remove noise from a generation until a plausible result is produced. Along the way however, inappropriate features like nudity or violence can be denoised, as seen in the training dataset. SLD updates the original prompt guidance encoding by computing the latent encoding of unsafe guidance, and shifting the original encoding away from the unsafe encoding given set hyperparameters. The resulting shifted "safe" guidance is far from unsafe embeddings, but still close enough to the original prompt guidance, to produce generations that attempt to follow the original prompt without inappropriate content. This augmented guidance is only provided after some warm-up steps of denoising to allow for the model to generate results close to the original prompt. The authors provide four sets of hyperparameters to configure the aggressiveness of the content filter. Using a model that detects inappropriate content, Schramowski et al. show that SLD has a lower probability of generating undesired features, and a user study supports

that the safe generations are still preferred the same as (or more than) the unaugmented results. It is important to note that, while SLD combats undesired generations, there are still instances of inappropriate features present due to their representation in the training dataset. Furthermore, the proposed algorithm can be misused by shifting the augmented guidance toward the unsafe vector, producing more inappropriate features.

## D Metrics and Evaluation Platforms

We now present common metrics, datasets, and simulation platforms leveraged when developing and evaluating the hallucination detection algorithms introduced in Section 4.

### D.1 Metrics

Here, we list established metrics used for computing language similarity and accuracy of generated image descriptions.

#### D.1.1 Language Similarity.

*BERTScore* [252]. Given a pair of responses, BERTScore computes the BERT [48] embeddings of the sentences and calculates their cosine similarity.

*BARTScore* [245]. Using a pre-trained BART model [112], which provides access to generated token probabilities, BARTScore sums over the log probability of each token generated while conditioning on context and previously output tokens. Essentially, BARTScore attempts to predict the quality of a generated text using BART as a proxy model.

*SummaC* [105]. SummaC is a class of natural language inference models that predict entailment, contradiction, and neutral scores between pairs of sentences among a document and its summary. Each score is collected into a separate matrix split by metric type. The authors propose two approaches, SummaC<sub>ZS</sub> and SummaC<sub>Conv</sub>, for aggregating scores of each sentence in the summary with respect to each sentence in the document.

*GPTScore* [65]. Like BARTScore, GPTScore relies on a pre-trained language model with access to token probabilities to estimate quality of outputs, but uses the GPT series of LLMs.

*AlignScore* [248]. The creators of AlignScore pose that two pieces of text are *aligned* when all information present in one text exists in the other, and the texts do not contradict one another. Consequently, they train a classification model on labeled data with three types of labels: a binary classification of aligned or not, a multi-class prediction including a neutral label in addition to the binary classification labels, and a continuous score for a regression task. The AlignScore metric computes a weighted score across all three prediction heads at test-time.

*Semantic Uncertainty* [101]. One common method of measuring uncertainty of a model’s many generations is computing its entropy over all generated token probabilities. However, in cases where multiple sentences have the same semantic meaning but output different entropies, the aggregated measurement is not representative of the true uncertainty of the model. Kuhn et al. tackle this problem by clustering sentences into semantic classes and summing entropies of sentences from the same class together.

#### D.1.2 Image Captioning.

*CHAIR* [175]. CHAIR<sub>i</sub>, used for measuring accuracy of descriptions of images, is the ratio of the number of hallucinated objects to all the objects mentioned in the description. To identify

the hallucinated objects within the description, the authors assume access to ground-truth object classes in the image.

*POPE* [119]. Li et al. recognize that different instructions prompting for a description of an image may lead to different responses from the model with the same semantic meaning. In this case, CHAIR gives different scores to both descriptions although they are alike. Instead, their proposed metric, POPE, asks binary questions about the existence of in-domain and out-of-domain objects in the image, which leads to more a more stable metric across different outputs.

## D.2 Offline Datasets

In this section, we present relevant offline datasets used for evaluating the performance of hallucination detection and mitigation techniques in driving, robotic, and QA tasks.

### D.2.1 Driving.

*BDD-X* [97]. BDD-X is a multi-modal driving dataset consisting of 20K samples (*i.e.*, video clips), each consisting of eight images with vehicle control actions and text annotations describing the scene and justifying actions.

*DriveGPT4* [231]. Xu et al. augment BDD-X into a QA dataset consisting of questions that ask about the current action of the vehicle, reasoning behind the action, and predicting future control signals. To incorporate other questions a user might ask about the vehicle, surroundings, and other miscellaneous queries, they prompt ChatGPT to generate further questions. In total, the DriveGPT4 dataset contains 56K samples.

*nuScenes* [22]. The nuScenes dataset contains 1K driving videos, each running for 20 seconds, collected from roads in Boston and Singapore. Each frame includes six different RGB camera views, GPS, annotated 3D bounding boxes of various object classes, and semantically labeled rader, lidar, and map representations.

*NuScenes-QA* [163]. Like DriveGPT4, NuScenes-QA is a visual QA dataset, but built on top of nuScenes. It includes five different types of questions including checking the existence of objects, counting instances, detecting the object being referred to, identifying the action state of an object, and comparing two objects. Overall, the dataset holds 450K QA pairs across 34K scenes in nuScenes.

*Talk2Car* [47]. Talk2Car is an earlier extension of the nuScenes dataset which aims to ignite further research into developing systems that bridge the gap between passengers and an autonomous vehicle through natural language. Annotators provided approximately 12K text commands over 850 videos within the nuScenes training split which refer to an object in the scene.

*Refer-KITTI* [223]. While Talk2Car is a pioneering work for object referral in real driving scenes through natural language, each annotated instruction only refers to one object. As such, Wu et al. propose a new task definition, referring multi-object tracking (RMOT), which attempts to predict all objects that are referred to within a natural language input. They augment the KITTI driving dataset [67] with labeled 2D bounding boxes around objects that are referenced within a text prompt for 6.5K images.

*NuPrompt* [224]. NuPrompt is another RMOT-based benchmark, but applied to nuScenes and with 3D bounding box labels. It includes 35K languages prompts, with most prompts referring to anywhere between one and ten objects.

*DRAMA* [135]. Malla et al. argue that, while several datasets exist for anomaly detection or identification on roads, there is a gap in explaining the reason for categorizing an object as being

risky, *i.e.*, objects the model should pay attention to, like crosswalks, pedestrians, and traffic lights. As such, DRAMA is a benchmark tackling identification of risky objects in a driving scene conditioned on natural language. Ding et al. [51] extend DRAMA to further include suggestions on actions the ego vehicle can take to minimize risk, but the dataset is not public at this time.

*NuInstruct* [50]. NuInstruct addresses two common limitations in existing driving datasets: they cover a limited subset of necessary tasks while driving (*e.g.*, evaluating perception while ignoring planning), and disregard temporal and multi-view representations. Built on top of NuScenes, the dataset provides 91K samples of multi-view sequences with corresponding QA pairs spanning 17 subtasks within perception, prediction, planning, and risk detection.

*DriveLM* [191]. The authors of DriveLM curate a similar comprehensive dataset from nuScenes and the CARLA driving simulator [54] with open-ended and factual questions about importance rankings of nearby vehicles, planning actions, detecting lanes, and more.

*Driving with LLMs* [29]. Chen et al. collect a text-based QA dataset from a proprietary driving simulator, generated from ChatGPT with ground-truth observations (*e.g.*, relative locations of detected vehicles, ego vehicle control actions, *etc.*) from the simulator.

### D.2.2 Code Generation and Robotics.

*HumanEval* [31]. HumanEval is a set of 164 handwritten programs, each with a function definition, docstring, program body, and unit tests. The authors find there is great promise in using LLMs for code generation, but output quality is limited by length of context and buggy examples.

*RoboCodeGen* [121]. Liang et al. build a new code generation benchmark specifically for robot tasks with 37 functions focused on spatial reasoning, geometric reasoning, and controls.

*Language-Table* [133]. The Language-Table dataset contains 594K trajectories manually annotated with 198K unique instructions across simulated and real-world manipulator robots. The multi-modal dataset consists of video sequences, corresponding actions at each time step, and language instructions describing the policy of the robot in hindsight.

*SaGC* [154]. The authors of the CLARA method developed a dataset to identify language goals from a user that are certain, ambiguous, and infeasible. Collected from three different types of robots (cooking, cleaning, and massage), SaGC is annotated with a floor-plan, descriptions of objects and people in view, a text goal, and a label of uncertainty.

### D.2.3 Question-answering.

*HotPotQA* [237]. HotPotQA is a question-answering benchmark with 113K multi-hop questions (*i.e.*, requiring multiple steps of reasoning to reach answer) collected from Wikipedia. The dataset includes both questions that require finding relevant phrases from context paragraphs, and comparing two entities.

*FEVER* [203]. In contrast to HotPotQA, the developers of FEVER attempt to answer the question of whether a fact is supported by a knowledge-base. The database contains 185K claims with annotated labels deciding if each claim is supported, refuted, or indeterminable from Wikipedia articles.

*Natural Questions* [103]. Natural Questions is yet another QA dataset with sources from Wikipedia. The authors release 307K training and 7K test samples of real (anonymized) queries into the Google search engine paired with a Wikipedia page and a long and short answer annotated by a person based on said article.

*StrategyQA* [68]. Like HotPotQA, StrategyQA aims to develop a dataset of implicit multi-hop questions, but includes a greater variety categories of questions, and with less category imbalance. Furthermore, most of the questions in the dataset require three or more steps of decomposition and referencing to accurately solve.

*QreCC* [5]. Separate from the information retrieval task described in benchmarks above, Anantha et al. develop a dataset, QreCC, for conversational QA. They focus on reading comprehension, passage retrieval, and question rewriting tasks, with a total of 13.7K dialogues paired with 81K questions.

*HA-DPO* [257]. Zhao et al. present a multi-model visual QA dataset of images, hallucinated descriptions, and non-hallucinated samples from the VG dataset [100].

#### D.2.4 Image and Video Generation.

*AVA* [145]. Murray et al. tackle curating a dataset to evaluate the aesthetic value of varying quality images. AVA contains 250K images with aesthetic ratings, textual semantic class labels, and pictographic style labels. The majority of images contain at least one semantic label, with 150K containing at least two labels.

*MSCOCO* [126]. The MSCOCO dataset contains 328K images with 2.5M human-labeled segmentations of common objects. Each image is accompanied by five text captions. In contrast to earlier datasets like ImageNet [46] and SUN [228], MSCOCO provides more instance labels per class and more non-iconic images containing multiple objects per scene.

*LSUN* [243]. As vision models continue to evolve and essentially *solve* existing benchmarks, new datasets need to be curated to provide additional challenges and evaluate generalization performance. Thus, Yu et al. propose a human-in-the-loop data labeling scheme to gather a dataset with one million images for ten scene and 20 object classes, by iterating between manual labeling, overfitting a classifier, and automatic labeling. Note that the overall precision of labels is slightly lower than pure manual labeling approaches.

*LAION-400M* [184]. More recently, companies are collecting proprietary datasets to train large vision models, outperforming open-sourced models trained on smaller scale datasets. To narrow the gap in training data for generative models, Schuhmann et al. present an image dataset with 400M images with corresponding metadata, CLIP [167] embeddings, and web crawling resources. Unfortunately, works have found that the training split of LAION-400M contains undesired content like racy imagery, slurs, stereotypes, and other biases, which should be filtered out [177].

*I2P* [182]. The Inappropriate Image Prompts dataset attempts to measure the likelihood of image generation models to output undesired, inappropriate content given a prompt. In particular, the dataset contains 4.5K prompts taken from real users of Stable Diffusion, which have been filtered to contain inappropriate details like hate speech, violent imagery, and criminal behavior.

*DrawBench* [177]. The DrawBench benchmark contains 200 prompts across 11 categories to evaluate the quality of generated images on a spectrum of properties (e.g., colors, quantities, generated text). Saharia et al. use their benchmark to present 8 samples from two different models to human raters, who choose which model they prefer qualitatively.

*T2VHaluBench* [36]. The developers of the Sora Detector curate a benchmark dataset for text to video hallucination detection. In total, the dataset contains 59, 8-second-long videos — 53 of which contain consistency, static, or dynamic hallucinations generated by the Runway-Gen-2 [3] and SORA [149] generative models.

### D.3 Simulation Platforms

Finally, we introduce common online simulators used to test hallucination detection methods for decision-making tasks.

#### D.3.1 Driving.

*HighwayEnv* [111]. Leurent presents a 2D car simulator, with driving scenarios ranging from a passing on a multi-lane highway, merging into a highway, merging and exiting from a roundabout, parking, and more. An ego vehicle can be controlled with discrete (*e.g.*, merge left, merge right, faster, *etc.*) or continuous (*e.g.*, providing an explicit acceleration command) actions.

*SUMO* [131]. Geared towards microscopic traffic simulation, SUMO allows researchers to design road networks, track traffic flow metrics, and control individual vehicles.

*CARLA* [54]. CARLA is a 3D driving simulator built on top of Unreal Engine. Existing works benchmark their methods on CARLA for perception, planning, control, and QA tasks for its realism. There is also capability to perform co-simulation with SUMO and CARLA simultaneously [216].

#### D.3.2 Robotics.

*Ravens* [246]. Ravens is a 3D manipulator robot (UR5e) simulator built with PyBullet [41] with tasks like block insertion, towers of hanoi, aligning boxes, assembling kits, *etc.* Each simulated task features a manipulator robot with a suction gripper sitting on a table workspace, with three camera views.

*ALFWorld* [190]. Building on top of the TextWorld simulator, discussed in Appendix D.3.3, ALF-World aligns perception from the 3D robot simulation benchmark, ALFRED [189], with text-based, discrete actions like “MoveAhead,” “RotateLeft,” and “Open.”

*ProgPrompt* [193]. ProgPrompt is a benchmark of high-fidelity 3D data collected from a virtual home robot. It includes three environments, each with 115 object instances. These simulations are further used to create a dataset of 70 household robot tasks with a ground-truth set of actions to achieve each goal.

*RoboEval* [80]. RoboEval is a general platform for checking the correctness of code generated for a robot task. It relies on a simulator, evaluator, and a set of defined tasks to perform evaluations on a simulated robot. While ProgPrompt captures more realistic scenarios in its high-fidelity 3D simulator, RoboEval is tuned towards verifying code efficiently.

*KnowNo TableSim* [174]. More recently, the developers of KnowNo also provide a tabletop simulator based on PyBullet, like Zeng et al. [246], for robot manipulation of blocks and bowls. Provided instructions vary in ambiguity by attribute, number, and spatial reasoning.

#### D.3.3 Other Simulators.

*TextWorld* [40]. TextWorld is a suite of text-based games that can be either hand-engineered or procedurally generated, where an agent directly receives text-based observations from an abstract world, and acts with natural language actions to complete a task.

*BabyAI* [34]. Chevalier-Boisvert et al. present a 2D top-down, grid-based simulator of instruction-following tasks with varying difficulty. Some tasks include simple navigation to a single goal, picking and placing objects with ambiguous references, and instructions that implicitly require multi-step reasoning to complete. The simulator provides a partial observation of the space near the agent at every timestep.

*MineDojo* [60]. The developers of MineDojo attempt to create a benchmark to test the continual learning of agents in an open-world setting. They build an interface on top of Minecraft, a video game, to enable testing with diverse open-ended tasks, and provide access to an external knowledge-base of existing Minecraft tutorials and wiki discussions. MineDojo includes several thousands of tasks that are more complex than earlier works (and require multi-step reasoning). As such, task completion is judged with a learned LVLM, which acts like a human evaluator.

*Smallville* [155]. Park et al. present a multi-agent conversational simulator where agents are controlled by language models. Users may set up agents with a defined backstory and provide instructions when desired. Each agent has access to a memory of past experiences, and generates natural language actions to go to certain areas, communicate with others, complete chores, and more.

## E Guidelines on Current Methodologies

### E.1 Beware of Erroneous Hallucination Predictions

We additionally point out risks of relying on hallucination intervention algorithms and metrics that use deep learning methods at inference time. Generally speaking, deep neural networks are black boxes with few statistical guarantees and lack interpretability [2, 14, 21, 52, 159, 201]. As such, these models may incorrectly detect hallucinations during deployment — effectively hallucinating themselves. In particular, engineers should take precautions when utilizing detection algorithms under method types like hidden states, attention weights, concept probabilities, analyzing samples, and proxy model, where neural networks are used frequently. Learned proxy models or classifiers are particularly prone to incorrect predictions because they are trained on a different data distribution from the model under test [8, 166]. Similarly, decreased model complexity may result in poor predictions [13, 79, 109]. As such, several works we have listed in Table 4 use LVLM-based evaluators (e.g., GPT-4) to increase evaluation model complexity and cover a broader data distribution. However, as we have found throughout this work, LVLMs should not be overly relied upon for consistently accurate estimates. Instead, learned hallucination detection models that further output a calibrated confidence score (relaying their uncertainty of a prediction) can assist designers with choosing how to utilize the evaluated model’s decision. Additionally, classic machine metrics like accuracy, precision, recall, false-positive rate (FPR), etc. provide engineers with an understanding of the effectiveness of hallucination detection methods prior to deployment. The learned intervention algorithm should be tuned to balance true-positive rate (TPR) and FPR, such that it does not miss critical hallucinations, nor act overly conservative (predicting hallucinations too frequently). We also suggest caution when using learning-based similarity metrics like BERTScore, SummaC, AlignScore, CLIP Score, and others listed in Table 4 for similar reasons. Specifically, while these metrics have been shown to reasonably reflect the semantic similarity of varied terms, there are still cases of poor alignment [71, 78]. Even adversarial models used in adversarial prompting approaches are not safe from learned biases, which could lead to a poor understanding of the uncertainty of the model under test. For example, an adversarial agent tasked with prompting an autonomous driving LVLM could provide incorrect sensor readings, obviously resulting in poor decisions from the LVLM under test. In this case, we have not learned any additional information on the reliability of the predicted action. Thus, we argue that deep-learning-based adversarial prompters should be grounded in accurate data to truly understand the uncertainty of model predictions. Overall, engineers should take care when using deep learning methods for hallucination intervention — or as metrics during evaluation — because of their tendency to act unpredictably with out-of-domain data, and limited theoretical guarantees.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Charu C. Aggarwal. 2023. *Neural Networks and Deep Learning*. Springer International Publishing, Cham.
- [3] Runway AI. 2023. Gen-2: Generate novel videos with text, images or video clips. *Runway blog* (2023). <https://runwayml.com/research/gen-2>
- [4] Soroush Ali, Glicksberg Benjamin S., Zimlichman Eyal, Barash Yiftach, Freeman Robert, Charney Alexander W., Nadkarni Girish N, and Klang Eyal. 2024. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI* 1, 5 (25 Apr 2024), 13 pages.
- [5] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Virtual)*. Association for Computational Linguistics, 520–534.
- [6] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv preprint arXiv:2110.01052* (2022).
- [7] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It’s Lying. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 967–976.
- [8] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. 2022. Agreement-on-the-line: Predicting the Performance of Neural Networks under Distribution Shift. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 19274–19289.
- [9] Song Bai and Jie Li. 2024. Progress and Prospects in 3D Generative AI: A Technical Overview including 3D human. *arXiv preprint arXiv:2401.02620* (2024).
- [10] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2404.18930* (2024).
- [11] Rounak Banik. 2017. The Movies Dataset. *Kaggle* (2017). <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>
- [12] Lauren Barack. 2024. Using AI in lesson planning? Beware hallucinations. *K-12 Dive* (2024). <https://www.k12dive.com/news/using-ai-lesson-planning-beware-hallucinations/726660/>
- [13] Falco J Bargagli Stoffi, Gustavo Cevolani, and Giorgio Grecco. 2022. Simple Models in Complex Worlds: Occam’s Razor and Statistical Learning Theory. *Minds and Machines* 32, 1 (March 2022), 13–42.
- [14] J.M. Benitez, J.L. Castro, and I. Requena. 1997. Are Artificial Neural Networks Black Boxes? *IEEE Transactions on Neural Networks* 8, 5 (1997), 1156–1164.
- [15] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving Image Generation with Better Captions. *OpenAI blog* (2023). <https://cdn.openai.com/papers/dall-e-3.pdf>
- [16] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. 2024. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. *arXiv preprint arXiv:2406.06512* (2024).
- [17] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2022. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2022).
- [18] Rakesh P Borase, DK Maghade, SY Sondkar, and SN Pawar. 2021. A review of PID control, tuning methods and applications. *International Journal of Dynamics and Control* 9 (2021), 818–827.
- [19] Neil Brewer, Rob Potter, Ronald P. Fisher, Nigel Bond, and Mary A. Luszcz. 1999. Beliefs and Data on the Relationship Between Consistency and Accuracy of Eyewitness Testimony. *Applied Cognitive Psychology* 13, 4 (1999), 297–313.
- [20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 2020 Conference on Neural Information Processing Systems (Virtual)*. Curran Associates, Inc., 1877–1901.
- [21] Nathan Buskulic, Jalal Fadili, and Yvain Quéau. 2024. Convergence and Recovery Guarantees of Unsupervised Neural Networks for Inverse Problems. *Journal of Mathematical Imaging and Vision* 66, 4 (Aug. 2024), 584–605.
- [22] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Virtual)*. Institute of Electrical and Electronics Engineers, 11618–11628.

- [23] Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin, Ireland). Association for Computational Linguistics, 6101–6119.
- [24] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the 2021 IEEE/CVF Conference on International Conference on Computer Vision (ICCV)* (Virtual). Institute of Electrical and Electronics Engineers, 9630–9640.
- [25] Neeloy Chakraborty, Aamir Hasan, Shuijing Liu, Tianchen Ji, Weihang Liang, D. Livingston McPherson, and Katherine Driggs-Campbell. 2023. Structural Attention-based Recurrent Variational Autoencoder for Highway Vehicle Anomaly Detection. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, 1125–1134.
- [26] Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions. *arXiv preprint arXiv:2305.14908* (2023).
- [27] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected?. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [28] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada). Association for Computational Linguistics, 1870–1879.
- [29] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)* (Yokohama, Japan). Institute of Electrical and Electronics Engineers, 14093–14100.
- [30] Liting Chen, Lu Wang, Hang Dong, Yali Du, Jie Yan, Fangkai Yang, Shuang Li, Pu Zhao, Si Qin, Saravan Rajmohan, et al. 2023. Introspective Tips: Large Language Model for In-Context Decision Making. *arXiv preprint arXiv:2305.11598* (2023).
- [31] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).
- [32] Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Zhibin Wang, Jingyi Yu, Gang Yu, et al. 2024. MeshXL: Neural Coordinate Field for Generative 3D Foundation Models. In *Proceedings of the 2024 Conference on Neural Information Processing Systems*.
- [33] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom). Association for Computing Machinery, New York, NY, USA, 245–255.
- [34] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. BabyAI: First Steps Towards Grounded Language Learning With a Human In the Loop. In *Proceedings of the 7th International Conference on Learning Representations* (New Orleans, LA, USA).
- [35] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research* 24, 240 (2023).
- [36] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. 2024. Sora Detector: A Unified Hallucination Detection for Large Text-to-Video Models. *arXiv preprint arXiv:2405.04180* (2024).
- [37] Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löfller, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine* 3, 1 (10 Oct 2023), 141.
- [38] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).
- [39] Tor Constantino. 2024. AI Experts Test Perplexity’s New Election Hub. *Forbes* (2024). <https://www.forbes.com/sites/torconstantino/2024/11/04/perplexitys-new-election-hub-triggers-reactions-from-ai-experts/>
- [40] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, et al. 2019. TextWorld: A Learning Environment for Text-Based Games. In *Proceedings of the 7th Computer Games Workshop at the 27th International Conference on Artificial Intelligence* (Stockholm, Sweden). Springer International Publishing, 41–75.

- [41] Erwin Coumans and Yunfei Bai. 2016–2021. PyBullet, a Python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>.
- [42] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A Survey on Multimodal Large Language Models for Autonomous Driving. In *Proceedings of the 1st Workshop on Large Language and Vision Models for Autonomous Driving at the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers, 958–979.
- [43] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (06 2024), 64–93.
- [44] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2136–2148.
- [45] Leonardo de Moura and Nikolaj Bjørner. 2008. Z3: An Efficient SMT Solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems* (Budapest, Hungary). Springer Berlin Heidelberg, 337–340.
- [46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Miami, FL, USA). Institute of Electrical and Electronics Engineers, 248–255.
- [47] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2Car: Taking Control of Your Self-Driving Car. 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)* (Hong Kong, China). Association for Computational Linguistics, 2088–2098.
- [48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, MN, USA). Association for Computational Linguistics, 4171–4186.
- [49] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason E Weston. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Proceedings of the Workshop on Reliable and Responsible Foundation Models at the 12th International Conference on Learning Representations* (Vienna, Austria).
- [50] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. 2024. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA). Institute of Electrical and Electronics Engineers, 13668–13677.
- [51] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. 2023. HiLM-D: Towards High-Resolution Understanding in Multimodal Large Language Models for Autonomous Driving. *arXiv preprint arXiv:2309.05186* (2023).
- [52] James E Dobson. 2023. On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities* 5, 2 (Nov. 2023), 431–449.
- [53] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. *arXiv preprint arXiv:2301.00234* (2023).
- [54] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Conference on Robot Learning* (Mountain View, CA, USA) (*Proceedings of Machine Learning Research*, Vol. 78). PMLR, 1–16.
- [55] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, HI, USA) (*Proceedings of Machine Learning Research*, Vol. 202). PMLR, 8469–8488.
- [56] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) (*Proceedings of Machine Learning Research*, Vol. 235). PMLR, 11733–11763.
- [57] L Ekenberg. 2000. The logic of conflicts between decision making agents. *Journal of Logic and Computation* 10, 4 (08 2000), 583–602.
- [58] Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and Reduction of Hallucinations in Open-Source Weak Large Language Models. *arXiv preprint arXiv:2308.11764* (2023).

- [59] Nidula Elgiriyewithana. 2023. Global Country Information Dataset 2023. *Kaggle* (2023). <https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023>
- [60] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 18343–18362.
- [61] Simone Formentin, Klaske van Heusden, and Alireza Karimi. 2013. Model-based and data-driven model-reference control: A comparative analysis. In *Proceedings of the 2013 European Control Conference (ECC)* (Zurich, Switzerland). Institute of Electrical and Electronics Engineers, 1410–1415.
- [62] Wikimedia Foundation. [n. d.]. *Wikimedia Downloads*. <https://dumps.wikimedia.org>
- [63] Markus Freitag and Yaser Al-Onaizan. 2017. Beam Search Strategies for Neural Machine Translation. In *Proceedings of the 1st Workshop on Neural Machine Translation* (Vancouver, Canada). Association for Computational Linguistics, 56–60.
- [64] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers, 910–919.
- [65] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Mexico City, Mexico). Association for Computational Linguistics, 6556–6576.
- [66] Jack Gallifant, Amelia Fiske, Yulia A. Levites Strekalova, Juan S. Osorio-Valencia, Rachael Parke, Rogers Mwavu, Nicole Martinez, Judy Wawira Gichoya, Marzyeh Ghassemi, Dina Demner-Fushman, et al. 2024. Peer review of GPT-4 technical report and systems card. *PLOS Digital Health* 3, 1 (01 2024), 1–15.
- [67] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI, USA). Institute of Electrical and Electronics Engineers, 3354–3361.
- [68] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* 9 (04 2021), 346–361.
- [69] Naresh Gupta and Dana S. Nau. 1992. On the complexity of blocks-world planning. *Artificial Intelligence* 56, 2 (1992), 223–254.
- [70] Junlin Han, Filippos Kokkinos, and Philip Torr. 2025. VFusion3D: Learning Scalable 3D Generative Models from Video Diffusion Models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)* (Milan, Italy). Springer Nature Switzerland, 333–350.
- [71] Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation (Virtual)*. Association for Computational Linguistics, Online, 507–517.
- [72] Frederick Hayes-Roth. 1985. Rule-based systems. *Commun. ACM* 28, 9 (sep 1985), 921–932.
- [73] Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. 2024. SayCanPay: Heuristic Planning with Large Language Models Using Learnable Domain Knowledge. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (Vancouver, Canada), Vol. 38. AAAI Press, 20123–20133.
- [74] Haoran He, Peilin Wu, Chenjia Bai, Hang Lai, Lingxiao Wang, Ling Pan, Xiaolin Hu, and Weinan Zhang. 2024. Bridging the Sim-to-Real Gap from the Information Bottleneck Perspective. In *8th Annual Conference on Robot Learning* (Munich, Germany).
- [75] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep Reinforcement Learning That Matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (New Orleans, LA, USA). AAAI Press, Article 392, 8 pages.
- [76] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the 9th International Conference on Learning Representations (Virtual)*.
- [77] Karl Moritz Hermann, Tomas Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 2015 Conference on Neural Information Processing Systems* (Montreal, Canada). Curran Associates, Inc.
- [78] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana, Dominican Republic). Association for Computational Linguistics, 7514–7528.
- [79] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. 2021. Model complexity of deep learning: a survey. *Knowledge and Information Systems* 63, 10 (Oct. 2021), 2585–2619.

- [80] Zichao Hu, Francesca Lucchetti, Claire Schlesinger, Yash Saxena, Anders Freeman, Sadanand Modak, Arjun Guha, and Joydeep Biswas. 2024. Deploying and Evaluating LLMs to Program Service Mobile Robots. *IEEE Robotics and Automation Letters* 9, 3 (2024), 2853–2860.
- [81] 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* (2023).
- [82] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA). Institute of Electrical and Electronics Engineers, 13418–13427.
- [83] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual Hallucinations of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.14683* (2024).
- [84] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716* (2024).
- [85] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA, USA). Institute of Electrical and Electronics Engineers, 6693–6702.
- [86] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, et al. 2023. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of The 6th Conference on Robot Learning* (Atlanta, GA, USA) (*Proceedings of Machine Learning Research*, Vol. 205). PMLR, 287–318.
- [87] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA). Institute of Electrical and Electronics Engineers, 857–866.
- [88] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. 2020. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends in Computer Graphics and Vision* 12, 1-3 (2020), 1–308.
- [89] Sumit Kumar Jha, Susmit Jha, Patrick Lincoln, Nathaniel D. Bastian, Alvaro Velasquez, Rickard Ewetz, and Sandeep Neema. 2023. Counterexample Guided Inductive Synthesis Using Large Language Models and Satisfiability Solving. In *Proceedings of the 2023 IEEE Military Communications Conference (MILCOM)* (Boston, MA, USA). Institute of Electrical and Electronics Engineers, 944–949.
- [90] 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. *Comput. Surveys* 55, 12, Article 248 (mar 2023), 38 pages.
- [91] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 1 (2019), 317.
- [92] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada). Association for Computational Linguistics, 1601–1611.
- [93] Haoqiang Kang and Xiao-Yang Liu. 2024. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*.
- [94] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. RealTime QA: What's the Answer Right Now?. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (New Orleans, LA, USA).
- [95] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring Catastrophic Forgetting in Neural Networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (New Orleans, LA, USA). AAAI Press, 3390–3398.
- [96] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. 2023. Can you text what is happening? Integrating pre-trained language encoders into trajectory prediction models for autonomous driving. *arXiv preprint arXiv:2309.05282* (2023).
- [97] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)* (Munich, Germany). Springer International Publishing, 577–593.
- [98] Seokhwan Kim, Spandana Gella, Chao Zhao, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Yang Liu, and Dilek Z Hakkani-Tur. 2023. Task-Oriented Conversational Modeling with Subjective Knowledge Track in DSTC11. In *Proceedings of The Eleventh Dialog System Technology Challenge* (Prague, Czech Republic). Association for Computational

- Linguistics, 274–281.
- [99] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. 2023. Segment Anything. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France). Institute of Electrical and Electronics Engineers, 3992–4003.
  - [100] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123 (2017), 32–73.
  - [101] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *Proceedings of the 11th International Conference on Learning Representations* (Kigali, Rwanda).
  - [102] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal Prediction with Large Language Models for Multi-Choice Question Answering. In *Proceedings of the ‘Neural Conversational AI Workshop - What’s left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) chatbots?’ at the 40th International Conference on Machine Learning* (Honolulu, HI, USA).
  - [103] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466.
  - [104] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward Design with Language Models. In *Proceedings of the The 11th International Conference on Learning Representations* (Kigali, Rwanda).
  - [105] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177.
  - [106] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2023. Large Language Models in Law: A Survey. *arXiv preprint arXiv:2312.03718* (2023).
  - [107] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, TX, USA, 1203–1213.
  - [108] Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, Cheap, and Powerful Refinement of LLMs. *arXiv preprint arXiv:2308.07317* (2023).
  - [109] Yoonho Lee, Juho Lee, Sung Ju Hwang, Eunho Yang, and Seungjin Choi. 2020. Neural Complexity Measures. In *Proceedings of the 2020 Conference on Neural Information Processing Systems* (Virtual). Curran Associates, Inc., 9713–9724.
  - [110] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6 (2015), 167–195. 2.
  - [111] Edouard Leurent. 2018. An Environment for Autonomous Driving Decision-Making. <https://github.com/eleurent/highway-env>.
  - [112] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Virtual). Association for Computational Linguistics, 7871–7880.
  - [113] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 2020 Conference on Neural Information Processing Systems* (Virtual). Curran Associates, Inc., 9459–9474.
  - [114] Chenghao Li, Chaoning Zhang, Joseph Cho, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. 2023. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. *arXiv preprint arXiv:2305.06131* (2023).
  - [115] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large Language Models with Controllable Working Memory. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics* (Toronto, Canada). Association for Computational Linguistics, 1774–1793.
  - [116] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A Dataset for Multi-Span Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, WA, USA). Association for Computational Linguistics, 1250–1260.

- [117] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 6449–6464.
- [118] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [119] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 292–305.
- [120] Yinhe Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (Brooklyn, NY, USA). Association for Computing Machinery, New York, NY, USA, 374–382.
- [121] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as Policies: Language Model Programs for Embodied Control. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)* (London, United Kingdom). Institute of Electrical and Electronics Engineers, 9493–9500.
- [122] Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. 2024. Introspective Planning: Guiding Language-Enabled Agents to Refine Their Own Uncertainty. *arXiv preprint arXiv:2402.06529* (2024).
- [123] Youngsun Lim and Hyunjung Shim. 2024. Addressing Image Hallucination in Text-to-Image Generation through Factual Image Retrieval. *arXiv preprint arXiv:2407.10683* (2024).
- [124] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research* (2022).
- [125] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin, Ireland). Association for Computational Linguistics, 3214–3252.
- [126] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)* (Zurich, Switzerland). Springer International Publishing, 740–755.
- [127] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).
- [128] Jiaqi Liu, Peng Hang, Xiao Qi, Jianqiang Wang, and Jian Sun. 2023. MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections. In *Proceedings of the 2023 IEEE International Conference on Intelligent Transportation Systems (ITSC)* (Bilbao, Spain). Institute of Electrical and Electronics Engineers, 5154–5161.
- [129] Jian Liu, Xiaoshui Huang, Tianyu Huang, Lu Chen, Yuenan Hou, Shixiang Tang, Ziwei Liu, Wanli Ouyang, Wangmeng Zuo, Junjun Jiang, and Xianming Liu. 2024. A Comprehensive Survey on 3D Content Generation. *arXiv preprint arXiv:2402.01166* (2024).
- [130] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Dublin, Ireland, 6723–6737.
- [131] Pablo Alvarez Lopez, Michael Behrisch, Laura Bicker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Weissner. 2018. Microscopic Traffic Simulation using SUMO. In *Proceedings of the 2018 International Conference on Intelligent Transportation Systems (ITSC)* (Maui, HI, USA). Institute of Electrical and Electronics Engineers, 2575–2582.
- [132] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 2507–2521.
- [133] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. 2023. Interactive Language: Talking to Robots in Real Time. *IEEE Robotics and Automation Letters* (2023).
- [134] Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2023. QUEST: A Retrieval Dataset of Entity-Seeking Queries with Implicit Set Operations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Toronto, Canada). Association for Computational Linguistics, 14032–14047.
- [135] Srikanth Malla, Chiho Choi, Ishit Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. DRAMA: Joint Risk Localization and Captioning in Driving. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision*

- (WACV) (Waikoloa, HI, USA). Institute of Electrical and Electronics Engineers, 1043–1052.
- [136] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 9004–9017.
- [137] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2024. A Language Agent for Autonomous Driving. In *Proceedings of the First Conference on Language Modeling*.
- [138] Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2023. FLIRT: Feedback Loop In-context Red Teaming. *arXiv preprint arXiv:2308.04265* (2023).
- [139] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA), Vol. 35. Curran Associates, Inc., 17359–17372.
- [140] Meta. 2024. Model Information. *GitHub* (2024). [https://github.com/meta-llama/llama-models/blob/main/models/llamas\\_1/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llamas_1/MODEL_CARD.md)
- [141] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Aslı Celikyilmaz, et al. 2023. Augmented Language Models: a Survey. *Transactions on Machine Learning Research* (2023). Survey Certification.
- [142] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FACTScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 12076–12100.
- [143] Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. 2022. DALL-E 2 Preview - Risks and Limitations. *GitHub* (2022). <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>
- [144] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [145] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI, USA). Institute of Electrical and Electronics Engineers, 2408–2415.
- [146] Jesutofummi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. 2024. Large Language Models in Medicine: The Potentials and Pitfalls. *Annals of Internal Medicine* 177, 2 (20 Feb 2024), 210–220.
- [147] OpenAI. 2022. Introducing ChatGPT. *OpenAI blog* (2022). <https://openai.com/blog/chatgpt>
- [148] OpenAI. 2023. DALL-E 3 System Card. *OpenAI blog* (2023). [https://cdn.openai.com/papers/DALL\\_E\\_3\\_System\\_Card.pdf](https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf)
- [149] OpenAI. 2024. Video generation models as world simulators. *OpenAI blog* (2024). <https://openai.com/index/video-generation-models-as-world-simulators/>
- [150] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaddin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024).
- [151] Gerhard Paaf and Sven Giesselbach. 2023. *Foundation Models for Speech, Images, Videos, and Control*. Springer International Publishing, Cham, 313–382.
- [152] Lorenz Pacchiardi, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Yue Pan, Yarin Gal, Owain Evans, and Jan M. Brauner. 2024. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [153] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, PA, USA). Association for Computational Linguistics, 311–318.
- [154] Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. 2024. CLARA: Classifying and Disambiguating User Commands for Reliable Interactive Robotic Agents. *IEEE Robotics and Automation Letters* 9, 2 (2024), 1059–1066.
- [155] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA). Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages.
- [156] Arkil Patel, Satwik Bhattacharya, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Virtual). Association for Computational Linguistics, 2080–2094.

- [157] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. *arXiv preprint arXiv:2302.12813* (2023).
- [158] Ethan Perez, Sam Ringer, Kamil Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics* (Toronto, Canada). Association for Computational Linguistics, 13387–13434.
- [159] Tomaso Poggio, Andrzej Banburski, and Qianli Liao. 2020. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30039–30045.
- [160] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *Proceedings of the 11th International Conference on Learning Representations* (Kigali, Rwanda).
- [161] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. VirtualHome: Simulating Household Activities Via Programs. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT, USA). Institute of Electrical and Electronics Engineers, 8494–8502.
- [162] Gokul Puthumanailam, Manav Vora, Pranay Thangeda, and Melkior Ornik. 2024. A Moral Imperative: The Need for Continual Superalignment of Large Language Models. *arXiv preprint arXiv:2403.14683* (2024).
- [163] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (Vancouver, Canada), Vol. 38. AAAI Press, 4542–4550.
- [164] Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. 2023. Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models. *arXiv preprint arXiv:2307.08487* (2023).
- [165] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal Language Modeling. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [166] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2008. *Dataset Shift in Machine Learning*. The MIT Press.
- [167] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Virtual) (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.
- [168] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019). <https://openai.com/research/better-language-models>
- [169] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX, USA). Association for Computational Linguistics, 2383–2392.
- [170] Anil Ramakrishna, Rahul Gupta, Jens Lehmann, and Morteza Ziyadi. 2023. INVITE: a Testbed of Automatically Generated Invalid Questions to Evaluate Large Language Models for Hallucinations. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore). Association for Computational Linguistics, 5422–5429.
- [171] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).
- [172] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. 2022. Red-Teaming the Stable Diffusion Safety Filter. In *NeurIPS ML Safety Workshop*.
- [173] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A Survey of Hallucination in Large Foundation Models. *arXiv preprint arXiv:2309.05922* (2023).
- [174] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. In *Proceedings of The 7th Conference on Robot Learning* (Atlanta, GA, USA) (*Proceedings of Machine Learning Research, Vol. 229*). PMLR, 661–682.
- [175] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium). Association for Computational Linguistics, 4035–4045.
- [176] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA). Institute of Electrical and Electronics Engineers, 10684–10695.
- [177] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic Text-to-Image Diffusion Models with

- Deep Language Understanding. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 36479–36494.
- [178] Pranab Sahoo, Prabhav Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Miami, FL, USA). Association for Computational Linguistics, 11709–11724.
- [179] Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino. 2021. Crossing the Reality Gap: A Survey on Sim-to-Real Transferability of Robot Controllers in Reinforcement Learning. *IEEE Access* 9 (2021), 153171–153187.
- [180] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA). Institute of Electrical and Electronics Engineers, 18582–18592.
- [181] Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, and Tianjie Zhu. 2024. Towards reducing hallucination in extracting information from financial reports using Large Language Models. In *Proceedings of the Third International Conference on AI-ML Systems* (Bangalore, India). Association for Computing Machinery, New York, NY, USA, Article 39, 5 pages.
- [182] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, Canada). Institute of Electrical and Electronics Engineers, 22522–22531.
- [183] Andre Schreiber, Tianchen Ji, D. Livingston McPherson, and Katherine Driggs-Campbell. 2023. An Attentional Recurrent Neural Network for Occlusion-Aware Proactive Anomaly Detection in Field Robot Navigation. In *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Detroit, MI, USA). Institute of Electrical and Electronics Engineers, 8038–8045.
- [184] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114* (2021).
- [185] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)* (Tel Aviv, Israel). Springer Nature Switzerland, 146–162.
- [186] Glenn Shafer and Vladimir Vovk. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* 9, 12 (2008), 371–421.
- [187] Dhruv Shah, Błażej Osiński, Brian Ichter, and Sergey Levine. 2023. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. In *Proceedings of The 6th Conference on Robot Learning* (Atlanta, GA, USA) (*Proceedings of Machine Learning Research*, Vol. 205). PMLR, 492–504.
- [188] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 2023 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 8634–8652.
- [189] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Virtual). Institute of Electrical and Electronics Engineers, 10737–10746.
- [190] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWORLD: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the 9th International Conference on Learning Representations* (Virtual).
- [191] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. DriveLM: Driving with Graph Visual Question Answering. *arXiv preprint arXiv:2312.14150* (2023).
- [192] Gonen Singer and Yuval Cohen. 2021. A framework for smart control using machine-learning modeling for processes with closed-loop control in Industry 4.0. *Engineering Applications of Artificial Intelligence* 102 (2021), 104236.
- [193] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)* (London, United Kingdom). Institute of Electrical and Electronics Engineers, 11523–11530.
- [194] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. 2024. LUNA: A Model-Based Universal Analysis Framework for Large Language Models. *IEEE Transactions on Software Engineering* 50, 7 (2024), 1921–1948.

- [195] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. 2023. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics* 3 (2023), 54–70.
- [196] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeib, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023).
- [197] Claire Stremple. 2024. False citations show Alaska education official relied on generative AI, raising broader questions. *Alaska Beacon* (2024). <https://alaskabeacon.com/2024/10/28/alaska-education-department-published-false-ai-generated-academic-citations-in-cell-policy-document/>
- [198] Naoki Suganuma and Keisuke Yoneda. 2022. Current Status and Issues of Traffic Light Recognition Technology in Autonomous Driving System. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E105.A, 5 (2022), 763–769.
- [199] Gaurav Suri, Lily R. Slater, Ali Ziaeef, and Morgan Nguyen. 2024. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. *Journal of Experimental Psychology: General* 153, 4 (04 2024), 1066–1075.
- [200] Gyana Swain. 2024. Patients may suffer from hallucinations of AI medical transcription tools. *CIO* (2024). <https://www.cio.com/article/3593403/patients-may-suffer-from-hallucinations-of-ai-medical-transcription-tools.html>
- [201] Mahsa Taheri, Fang Xie, and Johannes Lederer. 2021. Statistical guarantees for regularized neural networks. *Neural Networks* 142 (2021), 148–161.
- [202] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Virtual)*, Vol. 1.
- [203] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New Orleans, Louisiana). Association for Computational Linguistics, 809–819.
- [204] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (Istanbul, Turkey). European Language Resources Association, 2214–2218.
- [205] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860* (2024).
- [206] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).
- [207] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).
- [208] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *Proceedings of the 2017 International Semantic Web Conference* (Vienna, Austria). Springer International Publishing, 210–218.
- [209] Cem Uluoglakci and Tugba Temizel. 2024. HypoTermQA: Hypothetical Terms Dataset for Benchmarking Hallucination Tendency of LLMs. In *Proceedings of the Student Research Workshop at the 18th Conference of the European Chapter of the Association for Computational Linguistics* (St. Julian's, Malta). Association for Computational Linguistics, 95–136.
- [210] 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* (2023).
- [211] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85.
- [212] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (New Orleans, LA, USA).
- [213] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research* (2024).
- [214] Hongbo Wang, Jie Cao, Jin Liu, Xiaoqiang Zhou, Huabo Huang, and Ran He. 2024. Hallo3D: Multi-Modal Hallucination Detection and Mitigation for Consistent 3D Content Generation. In *Proceedings of the 2024 Conference on Neural Information Processing Systems*.

- [215] Jun Wang, Jiaming Tong, Kaiyuan Tan, Yevgeniy Vorobeychik, and Yiannis Kantaros. 2024. Conformal Temporal Logic Planning using Large Language Models. *arXiv preprint arXiv:2309.10092* (2024).
- [216] Axel Wegener, Michał Piórkowski, Maxim Raya, Horst Hellbrück, Stefan Fischer, and Jean-Pierre Hubaux. 2008. TraCI: an interface for coupling road traffic and network simulators. In *Proceedings of the 11th Communications and Networking Simulation Symposium* (Ottawa, Canada). Association for Computing Machinery, New York, NY, USA, 155–163.
- [217] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 2022 Conference on Neural Information Processing Systems* (New Orleans, LA, USA). Curran Associates, Inc., 24824–24837.
- [218] Benjamin Weiser. 2023. Here’s What Happens When Your Lawyer Uses ChatGPT. *The New York Times* (2023). <https://www.nytimes.com/2023/05/27/nyregion/aviaanca-airline-lawsuit-chatgpt.html>
- [219] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Association for Computational Linguistics, Copenhagen, Denmark, 94–106.
- [220] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao MA, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. 2024. DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [221] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao MA, Yingxuan Li, Linran XU, Dengke Shang, et al. 2024. On the Road with GPT-4V(ision): Explorations of Utilizing Visual-Language Model as Autonomous Driving Agent. In *Proceedings of the Workshop on Large Language Model (LLM) Agents at the 12th International Conference on Learning Representations* (Vienna, Austria).
- [222] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8, 3–4 (may 1992), 229–256.
- [223] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. 2023. Referring Multi-Object Tracking. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, Canada). Institute of Electrical and Electronics Engineers, 14633–14642.
- [224] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. 2023. Language Prompt for Autonomous Driving. *arXiv preprint arXiv:2309.04379* (2023).
- [225] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. *arXiv preprint arXiv:2402.02008* (2024).
- [226] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564* (2023).
- [227] Weihao Xia and Jing-Hao Xue. 2023. A Survey on Deep Generative 3D-aware Image Synthesis. *Comput. Surveys* 56, 4, Article 90 (nov 2023), 34 pages.
- [228] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA, USA). Institute of Electrical and Electronics Engineers, 3485–3492.
- [229] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [230] Yuanyu Xu, Zongxin Yang, and Yi Yang. 2023. SEEAvatar: Photorealistic Text-to-3D Avatar Generation with Constrained Geometry and Appearance. *arXiv preprint arXiv:2312.08889* (2023).
- [231] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. 2024. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters* 9, 10 (2024), 8186–8193.
- [232] Jianing Yang, Xuwei Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 2024. 3D-GRAND: A Million-Scale Dataset for 3D-LLMs with Better Grounding and Less Hallucination. *arXiv preprint arXiv:2406.05132* (2024).
- [233] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data* (feb 2024).
- [234] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for Honesty. In *Proceedings of the 2024 Conference on Neural Information Processing Systems* (Vancouver, Canada).
- [235] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli Vanderbilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. 2024. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In

- Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA). Institute of Electrical and Electronics Engineers, 16277–16287.
- [236] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2024. LLM4Drive: A Survey of Large Language Models for Autonomous Driving. In *Proceedings of the Workshop on Open-World Agents at the 2024 Conference on Neural Information Processing Systems* (Vancouver, Canada).
  - [237] ZhiLin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium). Association for Computational Linguistics, 2369–2380.
  - [238] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. *arXiv preprint arXiv:2310.01469* (2023).
  - [239] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations* (Kigali, Rwanda).
  - [240] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. *arXiv preprint arXiv:2309.06794* (2023).
  - [241] Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Bangkok, Thailand). Association for Computational Linguistics, 9333–9347.
  - [242] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D’Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog System Technology Challenge 7. *arXiv preprint arXiv:1901.03461* (2019).
  - [243] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-Scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365* (2015).
  - [244] Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2024. ReEval: Automatic Hallucination Evaluation for Retrieval-Augmented Large Language Models via Transferable Adversarial Attacks. In *Findings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (Mexico City, Mexico). Association for Computational Linguistics, 1333–1351.
  - [245] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Proceedings of the 2021 Conference on Neural Information Processing Systems* (Virtual). Curran Associates, Inc., 27263–27277.
  - [246] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. 2021. Transporter Networks: Rearranging the Visual World for Robotic Manipulation. In *Proceedings of the 4th Conference on Robot Learning* (London, United Kingdom) (*Proceedings of Machine Learning Research*, Vol. 155). PMLR, 726–747.
  - [247] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S. Yu. 2023. Large Language Models for Robotics: A Survey. *arXiv preprint arXiv:2311.07226* (2023).
  - [248] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Toronto, Canada). Association for Computational Linguistics, 11328–11348.
  - [249] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics* 3, 4 (2023), 100131.
  - [250] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How Language Model Hallucinations Can Snowball. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) (*Proceedings of Machine Learning Research*, Vol. 235). PMLR, 59670–59684.
  - [251] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2024. The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Bangkok, Thailand). Association for Computational Linguistics, 2025–2038.
  - [252] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations* (Virtual).
  - [253] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219* (2023).
  - [254] Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024. Revolutionizing Finance with LLMs: An Overview of Applications and Insights. *arXiv preprint arXiv:2401.11641* (2024).

- [255] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (Canberra, ACT, Australia). 737–744.
- [256] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023).
- [257] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. *arXiv preprint arXiv:2311.16839* (2024).
- [258] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *arXiv preprint arXiv:2302.09419* (2023).
- [259] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 12th International Conference on Learning Representations* (Vienna, Austria).
- [260] Cai-Nicolas Ziegler. 2020. Books Dataset. *Kaggle* (2020). <https://www.kaggle.com/datasets/saurabhbagchi/books-dataset>

Received 29 April 2024; revised 13 January 2025; accepted 3 February 2025