LightHouse: A Survey of AGI Hallucination

Feng Wang
Soochow University
zurichrain403@gmail.com

★ https://github.com/ZurichRain/AGI-Hallucination

Abstract

With the development of artificial intelligence, large-scale models have become increasingly intelligent. However, numerous studies indicate that hallucinations within these large models are a bottleneck hindering the development of AI research. In the pursuit of achieving strong artificial intelligence, a significant volume of research effort is being invested in the AGI (Artificial General Intelligence) hallucination research. Previous explorations have been conducted in researching hallucinations within LLMs (Large Language Models). As for multimodal AGI, research on hallucinations is still in an early stage. To further the progress of research in the domain of hallucinatory phenomena, we present a bird's eye view of hallucinations in AGI, summarizing the current work on AGI hallucinations and proposing some directions for future research. We will continuously update recent work on https://github.com/ZurichRain/AGI-Hallucination.

"Once this problem is solved, the path to AGI unfolds. We called it LightHouse for AGI."
—Anonymous

1 Introduction

The research in deep learning on AGI has witnessed explosive growth, particularly with the advent of LLMs like GPT4 (OpenAI, 2023), LLaMA (Touvron et al., 2023b), accelerating the arrival of the AGI era. LLMs have achieved remarkable results in many downstream natural language tasks. Simultaneously, the development of multimodal large models, such as LLaVA (Liu et al., 2023b), has sprung up like mushrooms after rain. There have also been outstanding achievements in the fields of vision, audio, 3D, and agent (Lin et al., 2023a; Gong et al., 2023; Hong et al., 2023; Szot et al., 2023a).

However, despite the astonishing progress in the LLM of AGI, the outputs of the model still do not fully align with human expectations, which called hallucination (Zhang et al., 2023c). For instance, LLMs occasionally produce factually incorrect answers, LVLMs(Large Vision-Language Models) can sometimes generate responses that object error or something of their own fantasy, video language models also suffer actions inconsistency, and 3D or agent may encounter issues with hallucination across multiple perspectives. This problems hinders AGI's path towards greater intelligence.

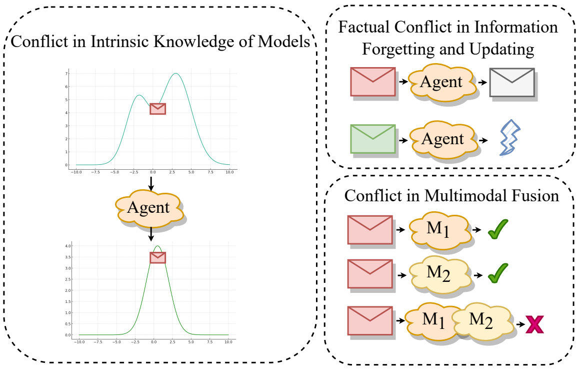

Conflict in Intrinsic Knowledge of Models
Agent
Factual Conflict in Information Forgetting and Updating
Agent
Agent
Conflict in Multimodal Fusion
M_1
M_2
M_1
M_2

Figure 1: Illustration of Hallucination in AGI. Which classify by three types: 1. Conflict in Intrinsic Knowledge of Models. 2. Factual Conflict in Information Forgetting and Updating. 3. Conflict in Multimodal Fusion. ✉ is input information, ✉ is forgotten information, ✉ is new information, and ⚡ is error output.

To mitigate the hallucinations in models during the AGI era, extensive works are underway. Currently, sparking in LLMs, RLHF (Christiano et al., 2017; Ouyang et al., 2022) (Reinforcement Learning from Human Feedback) is increasingly recognized as a prevalent method for mitigating hallucinatory phenomena in AI models. Enhances the model's responses to more closely align with human preferences through a reward-based RL mechanism. In LVLMs, LURE (Zhou et al., 2023b) used a hallucination revisor to reduce model's output hallucinations. There has a lot works engaged in fine-grained image captions to solve this problem. The 3D-LLM (Hong et al., 2023) employs 3D point clouds and additional 3D features as inputs, integrating the pre-trained knowledge of large models to uniformly address a variety of 3D-related tasks which include captioning, dense captioning, 3D question answering, task decomposition, 3D grounding, 3D-assisted dialogue, navigation et al. Additionally, significant efforts in the fields of video, audio, and agent domains have been undertaken to mitigate the hallucination inherent in models. They aslo used fine-grained mutimodel feature to improve model performance. Although hallucinations are not invariably detrimental, in certain instances, they can stimulate a model's creativity. Achieving equilibrium between hallucination and creative model output a significant challenge. Presently, there is an increase in research (Zhang, 2023; Yao et al., 2023) specifically talking about this problem.

Evaluating the hallucination error in AGI is also important area. By assessing LLMs, we gain deeper insights into their capabilities and limitations (Huang et al., 2023; Chiang and yi Lee, 2023; Chan et al., 2023a). Enhanced evaluation methodologies can significantly improve the interaction between humans and LLMs. This, in turn, has the potential to drive innovative designs and implementations in future interactions. Given the wide-ranging applications of LLMs, it is of utmost importance to ensure their safety and reliability, especially in sectors where safety is paramount, such as financial (Sarmah et al., 2023) institutions and healthcare facilities (Cai et al., 2023). There also has been a lot of progress in evaluating hallucinations with multimodal large models, including rule-based evaluation (Li et al., 2023f; Lovenia et al., 2023), large model-based evaluation (Wang et al., 2023a; Chan et al., 2023b), and human-based evaluation (Xu et al., 2023; Liu and Wan, 2023).

To improve clarity in understanding AGI hallucination within the AGI research and community and pave the way for the future research of AGI models, we are strongly motivated to compile this comprehensive survey. The structure of this article can be summarized in several parts as follows. First, we elucidated the concept of AGI Hallucination from three perspectives. Following this, we explore the emergence of AGI Hallucination. Then, we discuss the mitigation for AGI Hallucination within different domains. Finally, we talk about the evaluatation of AGI Hallucination, and discuss future research directions.

2 Definition for AGI Hallucination

Hallucination is a significant problem in AGI, research on hallucinations was predominantly derived from studies on LLMs in earlier stages. Recently, there has been an increasing focus on multimodal studies. In the early stages, researchers possessed an indistinct conceptualization of hallucinations, primarily focusing on the accuracy capabilities of models. Nowadays, we defined hallucinations as **model outputs that do not align with the contemporary empirical realities of our current world**. To further investigate hallucinations within AGI, we categorize them into the following three types:

- Conflict in Intrinsic Knowledge of Models
- Factual Conflict in Information Forgetting and Updating
- Conflict in Multimodal Fusion

2.1 Conflict in Intrinsic Knowledge of Models

Extensive research (Shen et al., 2021) indicates that there is a bias between the output distribution of the models and the distribution of the training data itself. Specifically, the model may produce hallucinations in the output due to learning biases. The investigation of such hallucinations in AGI is characterized as follows:

Language In LLMs, this type of hallucination often manifests as conflicts between model outputs and prompt, or inconsistencies within the context of the model outputs. As shown in Figure 4, the task is extracting all spatial relationships from a sentence, yet the model's response lacks many of the spatial relationships present in the input sentence which is a conflict between model outputs and model intrinsic knowledge.

Vision-Language Object Hallucination often occurs in LVLMs. For instance, in Figure 4, there is a computer next to the water cup in the image, but the model's response suggests that there is nothing beside the water cup. In some cases, there are also object errors. For example, the image shows 'a girl driving a motorcycle,' but the model's answer is 'the girl riding a horse'.

Video-Language Ullah and Mohanta (2022) shows that there are two kinds of hallucination in the Video-captioning: object and action hallucination. Object hallucination similar to Vision-Language. While action hallucination is shown in

Figure 4. The caption is "A woman is drawing a paper", and the ground truth is "A woman is folding a paper".

Audio-Language The model tends to use its language capability to answer the free-form open-ended question instead of conditioning on the audio input. And audio-language models also has a problem with “where is speaking part”. Figure 4 shows an error example that model confuses object recognition at the junction between text and audio.

3D-Language The hallucination of 3D is more complex than that of 2D, as it cannot be fully expressed from a single viewpoint, leading to inconsistencies in understanding across multiple perspectives. For example, in Figure 4, a 3D point cloud feature is provided along with two questions: "What does it look like from the front?" to which the model responds, "A horse." and "What does it look like from the left side?" to which the model responds, "A dog." . In this instance, the model exhibits inconsistency in its responses across multiple perspectives of the object.

2.2 Factual Conflict in Information Forgetting and Updating

Factual Conflict primarily arises when models fail to retain previously acquired factual knowledge and are unable to assimilate new information. Previous research on Factual Conflict has mainly focused on language models; recent research shows that the problem of Factual Conflict persists in multimodal contexts. In the following, we will explain the details of Factual Conflict exists in various modalities.

The hallucinations in the language models is frequently associated with the extent of the model’s underlying knowledge base (Augenstein et al., 2023; Melz, 2023). They also studied various scenarios in which models produce factual errors. Figure 5 also shows an example of this hallucination. In large image-text models, factual conflicts manifest in two ways. Firstly, the facts contained within the question may be influenced by an object or text in the image. For example in Figure 5, the text in image is "Barack Hussein Obama II", and the answer of text question is "Joseph Robinette Biden Jr.", while the model erroneously interprets the response to the inquiry as the text displayed in the image. Secondly, the recognition of objects in images may not align with real-world knowledge. In video and 3D tasks, factual inaccuracies are influenced not only by object recognition errors but also by inconsistency across multiple perspectives. Furthermore, in audio comprehension tasks, the model is easily influenced by homophones, leading to factual errors.

2.3 Conflict in Multimodal Fusion

Many existing multimodal large models integrate information from different modalities using an adapter method (Liu et al., 2023b). The fusion at the hidden state can easily induce errors from the pre-training stage of different modalities. Hallucination generated from multiple modalities can influence each other.

In the pre-training stage of image-text pairing, prevailing methodologies predominantly leverage contrastive learning or an Encoder-Decoder framework to derive image embeddings. While these approaches have yielded commendable outcomes in conventional image classification tasks, they are not without limitations, notably in terms of occasional inaccuracies in image recognition and a propensity for hallucinations in capturing intricate details of images. For example, the question is: Are badminton rackets usually placed together with shuttlecocks? However, in the picture, the badminton racket is placed together with a basketball, and the model incorrectly answered the question as: No. These include residual hallucinations in responses to image-related prompts, attributable to a partial comprehension of the visual content and potential biases ingrained in the alignment process.

In the realm of audio-language models, the issue of "where is speaking part" emerges as a notable challenge, which can be categorized under this type of hallucination. This issue arises when the model inadvertently overlooks the audio feature preceding the target audio feature. For example, in Figure 4, the content of the audio portion is: "Football plays a very important role in my life." while the entire model input is: "Here is a segment from an interview, [audio feature], what does the person being interviewed like?" The model occasionally confuse the features at the junction of text and audio during inference, which leads to the neglect or incorrect prediction of football in this example.

3 Emergence for AGI Hallucination

3.1 Training Data Distribution

The significance of training data on the efficacy of models is paramount. Both the quantity and quality of the data directly influence model’s distribution

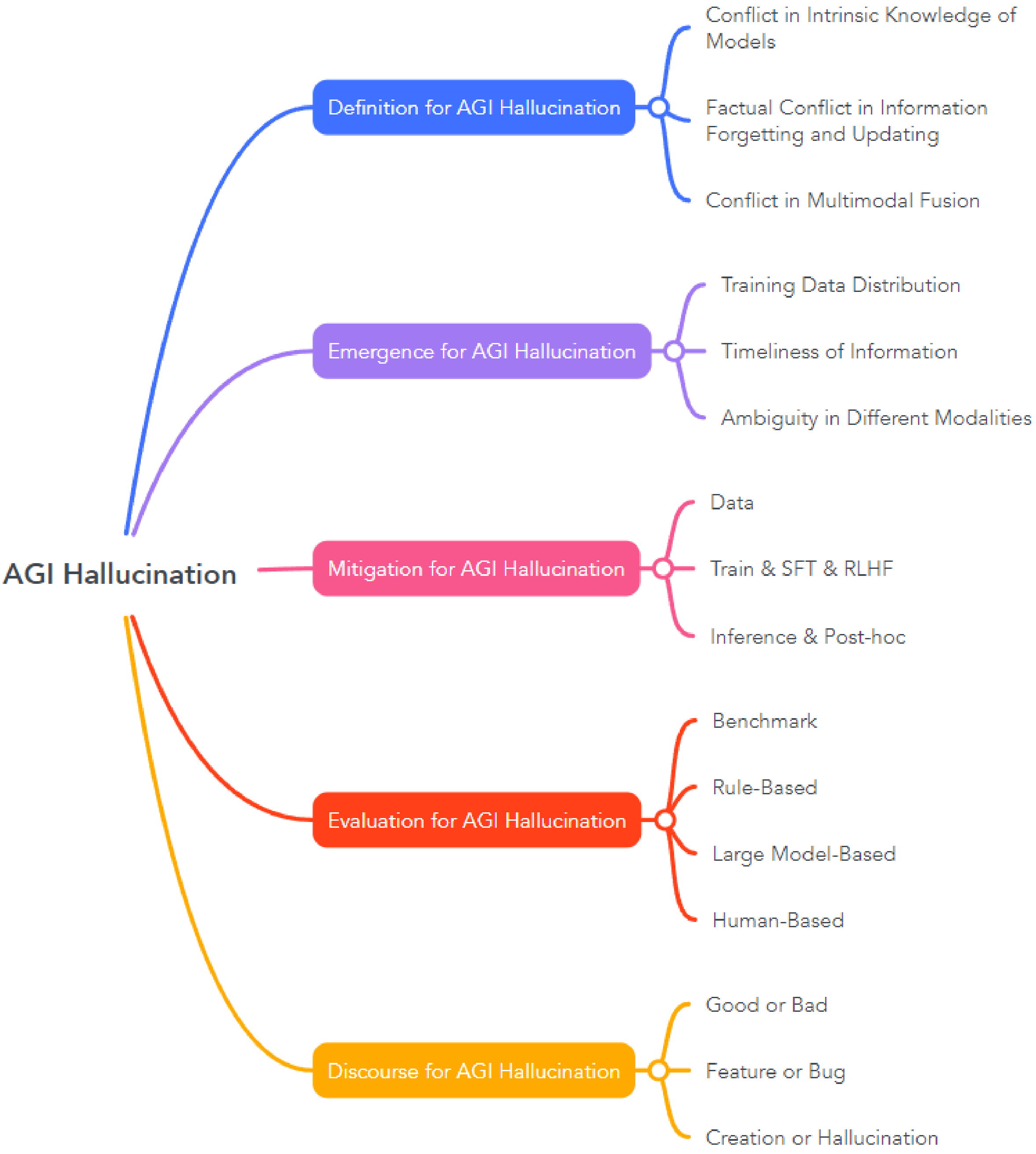

AGI Hallucination
Definition for AGI Hallucination
Conflict in Intrinsic Knowledge of Models
Factual Conflict in Information Forgetting and Updating
Conflict in Multimodal Fusion
Emergence for AGI Hallucination
Training Data Distribution
Timeliness of Information
Ambiguity in Different Modalities
Mitigation for AGI Hallucination
Data
Train & SFT & RLHF
Inference & Post-hoc
Evaluation for AGI Hallucination
Benchmark
Rule-Based
Large Model-Based
Human-Based
Discourse for AGI Hallucination
Good or Bad
Feature or Bug
Creation or Hallucination

Figure 2: The overview structure of this survey. we have explored the definition of hallucinations in the AGI, examined their causes, evaluated current works to mitigate them, assessed methods for hallucination analysis, engaged in dialogues concerning these phenomena, and contemplated future outlooks in the AGI.

(Hammoudeh and Lowd, 2022). The generalization ability of the model such as overfitting and underfitting to specific data distributions can adversely impact model function, occasionally leading to hallucinations in the model's outputs. In the machine learning, Mougan et al. (2023)'s research highlighted that shifts in training data distribution markedly alter explanation characteristics. This findings suggest that data distribution changes can profoundly influence a model's ultimate performance. Dong et al. (2023)'s work revealed that the quantity of compositional data significantly impacts model performance, whereas the ratio of composition bears minimal effect. Through integrating specialized domain data (mathematics and coding) into a general domain, the impact of varied data distributions on model performance was examined. The conclusion was that model capabilities are enhanced in scenarios with limited resources but diminish in high-resource contexts when compared to the capabilities derived from individual data sources. Furthermore, it was observed that as the model's size increases, there is a corresponding amplification in performance gains in low-resource settings, particularly for mathematical and general competencies. Dziri et al. (2022)'s findings indicate that the proportion of hallucinations in generated responses is substantially higher compared to that in the training data. LIMA (Zhou et al., 2023a) posit that the vast majority of knowledge in large language models is acquired during the pretraining phase, and that only a minimal amount of instruction tuning data is requisite for training these models to generate outputs of high quality. Li et al. (2020) stated that both overfitting and underfitting can impact the model's performance.

3.2 Timeliness of Information

Recent research underscores the critical importance of timeliness of information in preventing hallucinatory outputs from large models. Like humans, large models sometimes forget previous information. However, updating information in large models is quite challenging, leading to hallucinatory output due to outdated information. In the research, Lin et al. (2023b) has highlighted specific instances where foundational models, during the fine-tuning process, tend to sacrifice their general applicability in favor of becoming more specialized for particular tasks. Furthermore, Zhai et al. (2023b)'s work on evaluating catastrophic forgetting in Multimodel Large Language Models (MLLMs) is notable. By treating each MLLM as an image classifier, Zhai discovered that initial fine-tuning stages on an image dataset not only enhance performance across various other image datasets but also improve the alignment between textual and visual features. However, as fine-tuning continues, a significant drawback emerges: the MLLMs start to hallucinate. This leads to a considerable decline in their generalizability, a trend that persists even when the image encoder component remains unchanged.

3.3 Ambiguity in Different Modalities

The hallucinations that occur between multiple modalities are attributed to interactions among these modalities. Ambiguities between different modalities can lead to the model outputting incorrect knowledge. For instance, in the case of images with text, a conflict between the textual content in the image and the knowledge in the accompanying text can cause the model to output incorrect knowledge from the image. Comparable issues are also observable in 3D and video modalities. Wang et al. (2023b)'s research elucidates that hallucinations in LVLM primarily originate from discrepancies between visual and textual modalities.

4 Mitigation for AGI Hallucination

Due to the complexity of hallucinations and the black-box nature of neural networks, mitigating hallucinations is a difficult task. Current research in reducing hallucinations can be divided into different stages of the mitigation process, the data preparation (**Data**), the model training (**Train & SFT & RLHF**), and the model inference and post-processing (**Inference & Post-hoc**). We will detail the work of these three parts as follows:

Data In both the pre-training and fine-tuning stages, the distribution and quality of data are crucial. High-quality data usage during training can reduce model hallucinations. Ouyang et al. (2022)'s research shows that datasets with instructional annotations improve truthfulness compared to GPT-3 (Brown et al., 2020), signifying a major progression. They emphasized the importance of data distribution and quality in achieving robust model performance. Unlike previous approaches that handled each task individually, the data in Table 1 underscores the essential roles of tasks like Generation, OpenQA, and Brainstorming. Furthermore, the process of data cleansing is vital. For

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 1: The pre-training data distribution in (Ouyang et al., 2022).

instance, the creators of Llama 2 (Touvron et al., 2023c) intentionally upsample data from factual sources such as Wikipedia in constructing their pre-training corpus. Similarly, the developers of Falcon (Penedo et al., 2023) expertly extract high-quality data from the web using heuristic rules, underscoring that well-curated training corpora are foundational for effective LLMs. Wang et al. (2023c)'s study highlights the importance of fine-grained captions, which align more precisely with image representations, thus reducing hallucinations in multimodal contexts. Additionally, improving the resolution of images in datasets is shown to further mitigate hallucinations induced by the model. Li et al. (2023b) proposed using detailed video descriptions to train a VideoChat model. Moreover, the introduction of high-quality 3D features (point clouds and textures) or audio features (phonemes and tones) can alleviate the hallucinatory manifestations of the model.

Train & SFT & RLHF Numerous studies have demonstrated that applying suitable training techniques and loss functions during the training or supervise finetuning phase can significantly reduce model hallucination. LLaMA (Touvron et al., 2023a) reveals that instruction fine-tuning leads to rapid enhancements in Massive Multitask Language Understanding (MMLU) performance. LLaVA (Liu et al., 2023b) illustrates that a dual-stage training approach can effectively boost the efficacy of large multimodal models. In the first stage, the model synchronizes different modalities using image-caption datasets. Subsequently, in the second stage, the model is trained with multimodal instruction data. Yoon et al. (2022) developed the Text Hallucination Mitigating THAM framework, which integrates a THR (Text Hallucination Regularization) loss. This THR loss is derived from their novel information-theoretic approach to measuring text hallucination. Kanda et al. (2023) have developed an optimization scheme for ASR (Automatic Speech Recognition) focused on enhancing factual consistency, with the aim of reducing hallucinations of the ASR model.

Recent advancements increasingly indicate that RLHF (Reinforcement Learning from Human Feedback) is an efficacious approach to mitigate hallucinations in machine learning. RLHF comprises four sub-modules: reference model, base model, reward model, and value model. Specifically, GPT-4 (OpenAI, 2023) employs RLHF to enhance alignment with human preferences, significantly reducing instances of model hallucinations. Recentely, DPO (Rafailov et al., 2023) (Direct Preference Optimization) , has been proposed as a novel RLHF strategy. This method leverages just two models to effectively discern between positive and negative samples which further curtailing hallucinations. Sun et al. (2023b) introduces a novel alignment algorithm named Factually Augmented RLHF. This innovative approach enhances the reward model by incorporating additional factual data, including image captions and verified multi-choice options. This augmentation significantly mitigates the issue of reward hacking commonly encountered in RLHF, thereby elevating the algorithm's overall efficacy. In the context of agent-based systems, RL (Reinforcement Learning) is pivotal during both the training and inference phases. Emerging studies (Mu et al., 2022) suggest that employing multi-agent RL training methodologies can substantially enhance the task completion efficacy of these agents.

Inference & Post-hoc Post hoc explanation methodologies in machine learning are primarily categorized into perturbation-based and gradient-based approaches (Agarwal et al., 2021). These approaches are designed to modify input features during the model's inference phase, aligning the model's output more closely with human interpretative preferences. Perturbation-based methods involve constructing an interpretable surrogate model from the original, more opaque model. This is achieved through systematic perturbations of the input samples. Conversely, gradient-based methods, exemplified by techniques such as SmoothGrad (Smilkov et al., 2017) and Integrated Gradients, focus on computing the gradients of the model with

Area	Stage	Works
Language	*Data*	Brown et al. (2020), Ouyang et al. (2022), Touvron et al. (2023c), Penedo et al. (2023)
	Train & SFT & RLHF	Touvron et al. (2023a), OpenAI (2023), Rafailov et al. (2023)
	Inference & Post-hoc	Krishna et al. (2023), Cui et al. (2023), Feng et al. (2023a), Melz (2023), Zhao et al. (2021), Shapkin et al. (2023)
Visual Language	*Data*	Wang et al. (2023c)
	Train & SFT & RLHF	Liu et al. (2023b), Yoon et al. (2022), Sun et al. (2023b)
	Inference & Post-hoc	Agarwal et al. (2021), Peng et al. (2021), Zhao et al. (2021), Zhou et al. (2023b)
Video Language	*Data*	Li et al. (2023b)
	Train & SFT & RLHF	Li et al. (2023c), Jin et al. (2023a), Yu et al. (2023d), Zhang et al. (2023b), Shukor et al. (2023a), Jin et al. (2023b)
	Inference & Post-hoc	Jin et al. (2023c),Yin et al. (2023)
Audio Language	*Train & SFT & RLHF*	Kanda et al. (2023), Serai et al. (2022), Sridhar et al. (2023), Doh et al. (2023), Gong et al. (2023)
	Inference & Post-hoc	Lyu et al. (2023)
3D & Agent	*Train & SFT & RLHF*	Mu et al. (2022), Li et al. (2023e), Gong et al. (2022), Ren et al. (2023), Szot et al. (2023b), Xia et al. (2023)

Table 2: Overview of Migitation for AGI Hallucination.

respect to its input features. The process aids in determining the sensitivity of the model's output to each individual feature. In this context, Krishna et al. (2023) have developed an innovative technique In-Context Learning with Post Hoc Explanations. They automates the generation of rationales, effectively addressing the challenges presented by traditional post hoc explanation methods. LURE (Zhou et al., 2023b) uses the Revisor method to alleviate the LVLM hallucination by correcting the generated hallucination responses. krishna Sridhar et al. (2023) introduce a parameter-efficient, inference-time, faithful decoding algorithm that facilitates the use of compact audio captioning models. The results demonstrate performance on par with larger counterparts trained on extensive datasets.

Leveraging a knowledge-base plays a crucial role in mitigating hallucinations during model inference and post-processing. ChatLaw (Cui et al., 2023) proposes a strategy to counteract hallucination by augmenting the training process of the model and integrating four distinct modules during inference, namely 'consult', 'reference', 'self-suggestion', and 'response'. Feng et al. (2023a) endeavor to instruct LLMs in querying relevant domain-specific knowledge from external knowledge graphs for answering specialized queries. Melz (2023) employs RAG (Retrieval Augmented Generation), (Lewis et al., 2020), to enhance problem-solving efficacy. The Verify-then-Edit methodology (Zhao et al., 2021) is focused on improving the factuality of predictions through post-editing reasoning chains, using external information sourced from Wikipedia. Peng et al. (2021) adopts the information retrieval paradigm, automating the identification of related entities in text-image pairs to boost model performance. Zhao et al. (2021) introduces an innovative approach that constructs a MMKG (multi-modal knowledge graph), linking visual objects with named entities and simultaneously capturing the relationships among these entities, aided by external knowledge obtained from the web. Jin et al. (2023c) utilizes ClipBERT for extracting video-question features and derives frame-wise object-level external knowledge from a commonsense database to augment the performance of the model. Shapkin et al. (2023) introduces an end-to-end trainable architecture that incorporates a scalable entity retriever directly into the LLM decoder. This method significantly enhances the performance of code generation.

5 Evaluation for AGI Hallucination

The evaluation of AGI hallucinations is critically important, with substantial progress being made in AGI assessment in current research. Human evaluation of model-generated hallucinations is a time-consuming process, also humans themselves are susceptible to hallucinations. In the field of LLMs, Yu et al. (2023b); Sun et al. (2023a) have proposed the use of rule-based methods to detect model hallucinations, while Zha et al. (2023); Min et al. (2023); Liu and Wan (2023); Dhuliawala et al. (2023) have introduced automated assessment methods for detecting factual hallucinations, both of which are commendable contributions in automatically evaluate large model hallucinations. The occurrence of hallucinations in multimodal AGI is even more complex, encompassing not only misconceptions arising from language comprehension errors but also hallucinations related to the recognition of information in other modalities. In the context of LVLMs, POPE (Li et al., 2023f) first proposed the use of discriminative methods to detect object hallucinations in images. Additionally, the work of (Zhang et al., 2019; Yuan et al., 2021; Bai et al., 2023) highlights that methods based on LLMs can more effectively detect hallucinations. External knowledge bases also play a vital role in hallucination detection. In this section, we first introduce the benchmark for hallucination evaluation, then we categorize hallucination assessment into three types: rule-based, large model-based, and human-based.

Benchmark In both the traditional deep learning era and the era of large models, benchmarks play a crucial role in evaluating model illusions. Benchmarks in different fields, such as ChainPoll (Friel and Sanyal, 2023) and Bang et al. (2023), focus on assessing specific domain weaknesses, while benchmarks across different modalities evaluate various model skills. KoLA (Yu et al., 2023b) introduced a knowledge-driven benchmark comprising 19 tasks, utilizing Wikipedia data along with continuously collected emerging corpora to construct more granular evaluations, addressing unseen data and evolving knowledge. UHGEval (Liang et al., 2023) is an Unconstrained Hallucination Generation Evaluation benchmark for evaluating prominent Chinese language models. INVITE (Ramakrishna et al., 2023) automatically generates invalid questions to assess large model illusions. Meanwhile, Head-to-Tail (Sun et al., 2023a) uses a template-based method to automatically generate 18K QA (question-answer) pairs for model illusion assessment. DELUCIONQA (Sadat et al., 2023) employs a QA system to automatically generate domain-specific questions and answers for detecting model illusions. Numerous benchmarks have also emerged in the fields of image and video, particularly for LVLMs such as MMBench (Liu et al., 2023d), MM-Vet (Yu et al., 2023c), SEED-Bench (Li et al., 2023a), LVLM-eHub (Xu et al., 2023), and AutoEval-Video (Chen et al., 2023a). MERLIM (Villa et al., 2023) is an extensive database with over 279,000 image-question pairs, primarily focusing on identifying and analyzing cross-modal 'hallucination' events in IT-LVLMs (Image-Text Language-Vision Language Models). In multimodal contexts, finer-grained captions can detect more nuanced hallucinations in image-text interactions, Gunjal et al. (2023) with a dataset of 16k finely annotated VQA examples. HALLUSIONBENCH (Liu et al., 2023a) is the first to consider visual illusion and knowledge hallucination of LVLMs. RSGPT (Hu et al., 2023) introduces a benchmark concerning Remote Sensing. In the video domain, benchmarks like AutoEval-Video (Chen et al., 2023a) , EgoSchema (Mangalam et al., 2023), EvalCrafter (Liu et al., 2023c), and MVBench (Li et al., 2023d) have emerged. AutoEval-Video and MVBench establish video QA datasets, assessing Multi-modal Video from multiple perspectives. EgoSchema focuses on Very Long-form Video Language Understanding, while EvalCrafter and LLM4VG (Feng et al., 2023b) emphasizes the assessment of Video Generation. Himakunthala et al. (2023) evaluates complex video reasoning tasks, and Liu and Wan (2023) evaluates the factuality datasets for video captioning. Grauman et al. (2023) present an exocentric video of skilled human activities as multimodal multiview

Area	Eval-Mod	Works
Language	*Benchmark*	Friel and Sanyal (2023), Bang et al. (2023), Yu et al. (2023b), Liang et al. (2023), Ramakrishna et al. (2023),Sun et al. (2023a),Sadat et al. (2023)
	Rule-Based	Lin (2004), Papineni et al. (2002), Sun et al. (2023a), Min et al. (2023), Yu et al. (2023b)
	Large Model-Based	Zhang et al. (2019), Yuan et al. (2021), Zha et al. (2023), Chan et al. (2023b),Sarmah et al. (2023), Chan et al. (2023a)
	Human-Based	Chiang and yi Lee (2023), Min et al. (2023)
Visual Language	*Benchmark*	Bang et al. (2023), Liu et al. (2023d), Yu et al. (2023c), Li et al. (2023a), Xu et al. (2023),Villa et al. (2023), Gunjal et al. (2023), Liu et al. (2023a), Hu et al. (2023)
	Rule-Based	Shukor et al. (2023b), Vedantam et al. (2014), Rohrbach et al. (2018), Li et al. (2023f), Lovenia et al. (2023)
	Large Model-Based	Zhai et al. (2023a), Bai et al. (2023)
	Human-Based	Guan et al. (2023), Xu et al. (2023), fei Yin et al. (2023)
Video Language	*Benchmark*	Bang et al. (2023), Chen et al. (2023a), Chen et al. (2023a), Mangalam et al. (2023), Liu et al. (2023c), Li et al. (2023d), (Feng et al., 2023b), Himakunthala et al. (2023), Liu and Wan (2023), Grauman et al. (2023)
	Rule-Based	Vedantam et al. (2014), Rohrbach et al. (2018)
	Large Model-Based	Zhang et al. (2023a), Feng et al. (2023b), Chen et al. (2023b)
	Human-Based	Ma et al. (2023), Chen et al. (2023c)
Audio Language	*Benchmark*	Behera et al. (2023), de Seyssel et al. (2023)
	Rule-Based	Yu et al. (2023a)
3D & Agent	*Benchmark*	Behera et al. (2023), Chen et al. (2023c)
	Large Model-Based	Yang et al. (2023)
	Human-Based	fei Yin et al. (2023)

Table 3: Overview of Evaluation for AGI Hallucination.

video dataset and benchmark challenge. Behera et al. (2023) used LLM to generate expansive, high-quality Audio Question Answering datasets, contributing significantly to the progression of Audio research.

Rule-Based Rule-based evaluation methodologies have significantly evolved in traditional deep learning tasks, exemplified by the development of ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). However, the challenge intensifies when assessing hallucinations in single sentences, particularly in the context of LLMs. In the evaluation of LLMs, Head-to-Tail (Sun et al., 2023a) , employing a triad of metrics: accuracy (A), hallucination rate (H), and missing rate (M), to gauge the quality of the generated sentences. This approach further integrates exact match (EM), token F1 (F1), and ROUGE-L into its assessment criteria. Min et al. (2023), firstly, breaking a generation into a series of atomic facts—short statements, then assigning a binary label to each atomic fact, and finally evaluating the FACTSCORE in the sentence.

Furthermore, KoLA (Yu et al., 2023b) initiative proposed the innovative Self-contrast Metric, rooted in ROUGE-L. This metric computes an average contrast score by juxtaposing results derived from knowledge-based prompts, standard prompts, and human responses, thereby effectively evaluating hallucinations in LLM-generated content.

For LVLMs, Shukor et al. (2023b) adopted a unique approach, utilizing VQA accuracy, CIDEr (Vedantam et al., 2014), and CHAIRs (Rohrbach et al., 2018), specifically for appraising hallucinations within the COCO dataset. Object Hallucination, in particular, represents a formidable challenge in LVLMs. To address this problem, POPE (Li et al., 2023f) , based on the co-occurrence principle of objects in images. POPE, a polling-based query methodology, is designed to scrutinize model hallucinations on objects through binary "is or not" questions. In a parallel development, NOPE (Lovenia et al., 2023) leveraging LLMs to generate data featuring NegP (Negative Pronouns). NOPE's approach is instrumental in evaluating object hallucination in these models.

Large Model-Based The surge in popularity of Large Model-Based evaluation techniques is quite remarkable, particularly in contrast to Rule-Based methods. These Model-Based approaches excel in executing detailed assessments of models. BERTSCORE (Zhang et al., 2019) employs the BERT model to evaluate the similarity between candidate and reference sentences. BARTSCORE (Yuan et al., 2021) assesses model-generated sentences by calculating weighted marginal probability distributions, utilizing token probability distributions at each timestep. ALIGNSCORE (Zha et al., 2023) establishes a novel text-to-text information alignment function for assessing the factual consistency between two text pieces. CLAIR (Chan et al., 2023b) leverages the capabilities of highly credible LLMs like ChatGPT to generate both candidate scores and an explanatory rationale for the score. In the financial reporting sector, Sarmah et al. (2023) integrates BERTScore, BARTScore, Jaro similarity, and the Longest Common Subsequence method for evaluating model-generated reports. ChatEval (Chan et al., 2023a) employs a team of multi-agent referees to assess the response quality produced by various models in response to open-ended inquiries and conventional tasks in the domain of NLG (Natural Language Generation). HALLE-SWITCH (Zhai et al., 2023a) proposes CCEval, an innovative GPT-4 assisted evaluation tool for detailed captioning. Finally, TouchStone (Bai et al., 2023) employs GPT-4, a formidable language model, to assess the diverse capabilities of LVLMs. Kabra et al. (2023) conducted an evaluation of VLMs for 3D Objects, focusing on their responses to variations in object view, the phrasing of questions, previously inferred information specified in the prompt, and the accessibility of the object's visual characteristics.

Human-Based Human-based evaluation in assessing the precision and authenticity of models, particularly in aligning with human preferences. Chiang and yi Lee (2023) emphasizes the use of both human evaluation and LLM evaluation in two specific NLP tasks: open-ended story generation and adversarial attacks. The findings indicate that evaluations conducted by LLMs can lead to ethical concerns.

Min et al. (2023) involves human annotators assigning one of three labels (Irrelevant, Supported, Not-supported) to each atomic fact, enhancing the precision of the assessment. In contrast, Guan et al. (2023) entails human experts manually collecting 346 images across a variety of topics and types. These experts not only select each image meticulously but also compose corresponding question-answer pairs.

Liu and Wan (2023) has developed an annotation protocol aimed at guiding annotators in evaluating and labeling the factuality of video captions. This method led to the creation of two human-annotated factuality datasets: ActivityNet-Fact, comprising 200 videos and 3,152 sentences, and YouCook2-Fact, including 100 videos and 3,400 sentences.

Additionally, the passage mentions LVLM-eHub (Xu et al., 2023) and LAMM (fei Yin et al., 2023), which utilize existing public datasets from various computer vision tasks for evaluation purposes. These evaluations are conducted either by human annotators or GPT models. LAMM, in particular, extends its scope to encompass a broad range of vision tasks in both 2D and 3D domains.

6 Discourse for AGI Hallucination

Mitigating hallucinations is essential in AGI models, it is also important to notice that not all such occurrences are detrimental. In some scenarios, hallucinations can induce the model's creativity. Striking a balance between hallucination and creation is a crucial challenge.

Recent work has also shown that hallucinations are not entirely erroneous. Hallucinations play a role as adversarial examples to increase the robustness and creation of models, which need to produce hallucinations in contextually reasonable situations - akin to 'white lie' in human life. Yao et al. (2023)'s research utilized weak semantic prompts and OOD prompts to elicit hallucinatory responses from LLMs. This approach led to a reassessment of hallucinations as a different perspective on adversarial examples. Zhang (2023) focused on the degree of faithfulness to reference knowledge in generated responses, striving for a balance between creativity and hallucination. Zhao et al. (2023) employed a method called Hallucination-Aware Direct Preference Optimization, using hallucinatory samples to optimize models through reinforcement learning, demonstrating this method's effectiveness in mitigating hallucinations in LVLMs. Qiu et al. (2022) introduced an approach where a teacher iteratively generates synthetic training data based on the learner's status, a process termed data hallucination teaching. Wu et al. (2023) developed a 'Hallucinator' to generate additional positive image samples for enhanced contrast training. Fei et al. (2023) conceived a visual scene hallucination mechanism that dynamically creates pseudo visual scene graphs from textual scene graphs, significantly improving inference-time image-free unsupervised multimodal machine translation. Kulal et al. (2023) proposed inserting characters into scenes, enabling models to generate videos with both character and scene hallucinations, achieving harmonious composition and creativity. McKee et al. (2021) initially added video simulation augmentations to create hallucinated video data and then trained a tracker jointly on this hallucinated data and mined hard video examples. HALLUAUDIO (Yu et al., 2023e) leverages a special audio format by hallucinating high-frequency and low-frequency parts as structured concepts for few-shot audio classification. Finally, Shah et al. (2023) pioneered the use of hallucinated latent positives in a skeleton-based CL (Contrastive Learning) framework. They also employed a MoCo-based framework that mixes latent space features of positives and negatives, better utilizing hallucinations in action generation tasks.

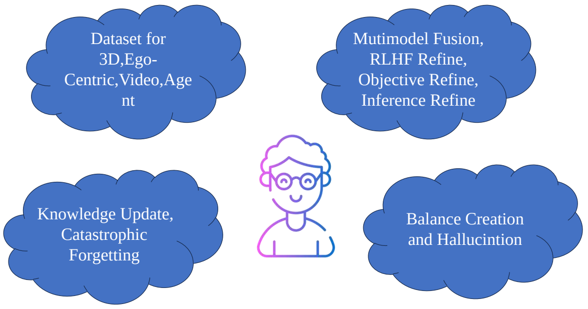

Dataset for 3D,Ego-Centric,Video,Age nt
Mutimodel Fusion, RLHF Refine, Objective Refine, Inference Refine
Knowledge Update, Catastrophic Forgetting
Balance Creation and Hallucintion

Figure 3: Talking about Future.

7 Talking about Future

The model generate hallucinations is influenced by data, model architectures, and information updating. Figure 3 shows the future about AGI (Hallucination) Research. Presently, there's a notable deficit of high-quality data in the spheres of audio, 3D modeling, and agent-based systems. Future studies must prioritize the development of robust datasets in these areas. It is notice whether tokens are suitable representations for modalities that non-textual context and how models across different modalities can be better integrated. For knowledge updates, it's imperative for models to emulate human-like short-term and long-term memory functions. Investigating methods to enhance knowledge updating in models while retaining their foundational information is a critical area of research. Additionally, the equilibrium between hallucination and creation is crucial. Investigating methodologies to enable models to appropriately engage in hallucinatory activities presents a significant and interesting research avenue.

Dispel the clouds of hallucination around the AGI, and build a true AGI.

References

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Zhiwei Steven Wu, and Himabindu Lakkaraju. 2021. Towards the unification and robustness of perturbation and gradient based explanations. In *International Conference on Machine Learning*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram A. Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *ArXiv*, abs/2310.05189.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xing Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *ArXiv*, abs/2308.16890.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Swarup Ranjan Behera, Krishna Mohan Injeti, Jaya Sai Kiran Patibandla, Praveen Kumar Pokala, and Balakrishna Reddy Pailla. 2023. Aquallm: Audio question answering data generation using large language models. *arXiv preprint arXiv:2312.17343*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. Medbench: A large-scale chinese benchmark for evaluating medical large language models. *arXiv preprint arXiv:2312.12806*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023a. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201.

David Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John F. Canny. 2023b. Clair: Evaluating image captions with large language models. *ArXiv*, abs/2310.12971.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2023a. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *ArXiv*, abs/2311.14906.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2023b. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2023c. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *arXiv preprint arXiv:2312.06722*.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics*.

Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741.

Jiaxi Cui, Zongjia Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *ArXiv*, abs/2306.16092.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2023. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *arXiv preprint arXiv:2312.14069*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv*, abs/2309.11495.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *ArXiv*, abs/2310.05492.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *North American Chapter of the Association for Computational Linguistics*.

Hao Fei, Qianfeng Liu, Meishan Zhang, M. Zhang, and Tat seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Annual Meeting of the Association for Computational Linguistics*.

Zhen fei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Wanli Ouyang, and Jing Shao. 2023. Lamm: Language-assisted multimodal instruction-tuning dataset, framework, and benchmark. *ArXiv*, abs/2306.06687.

Chao Feng, Xinyu Zhang, and Zichu Fei. 2023a. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *ArXiv*, abs/2309.03118.

Wei Feng, Xin Wang, Hong Chen, Zeyang Zhang, Zihan Song, Yuwei Zhou, and Wenwu Zhu. 2023b. Llm4vg: Large language models evaluation for video grounding. *arXiv preprint arXiv:2312.14206*.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *ArXiv*, abs/2310.18344.

Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, and Xinchao Wang. 2022. Posetriplet: Co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11017–11027.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *ArXiv*, abs/2305.10790.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2023. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu Ruiqi Xian Zongxia Li, Xiaoyu Liu Xijun Wang, Lichang Chen Furong Huang Yaser Yacoob, and Dinesh Manocha Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310.

Anish Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *ArXiv*, abs/2308.06394.

Zayd Hammoudeh and Daniel Lowd. 2022. Training data influence analysis and estimation: A survey. *arXiv preprint arXiv:2212.04612*.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Philip Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Stephen Saxon, and William Yang Wang. 2023. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Conference on Empirical Methods in Natural Language Processing*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *ArXiv*, abs/2307.12981.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. 2023. Rsgpt: A remote sensing vision language model and benchmark. *ArXiv*, abs/2307.15266.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023a. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Yao Jin, Guocheng Niu, Xinyan Xiao, Jian Zhang, Xi Peng, and Jun Yu. 2023b. Knowledge-constrained answer generation for open-ended video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8141–8149.

Yao Jin, Guocheng Niu, Xinyan Xiao, Jian Zhang, Xi Peng, and Jun Yu. 2023c. Knowledge-constrained answer generation for open-ended video question answering. In *AAAI Conference on Artificial Intelligence*.

Rishabh Kabra, Loïc Matthey, Alexander Lerchner, and Niloy Jyoti Mitra. 2023. Evaluating vlms for score-based, multi-probe annotation of 3d objects. *ArXiv*, abs/2311.17851.

Naoyuki Kanda, Takuya Yoshioka, and Yang Liu. 2023. Factual consistency oriented speech recognition. *ArXiv*, abs/2302.12369.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. *ArXiv*, abs/2305.11426.

Arvind krishna Sridhar, Yinyi Guo, Erik Visser, and Rehana Mahfuz. 2023. Parameter efficient audio captioning with faithful guidance using audio-text shared latent representation. *ArXiv*, abs/2309.03340.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. 2023. Putting people in their place: Affordance-aware human insertion into scenes. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17089–17099.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2023d. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv*, abs/2311.17005.

Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. 2023e. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.10763*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023f. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.

Zichao Li, Liyuan Liu, Chengyu Dong, and Jingbo Shang. 2020. Overfitting or underfitting? understand robustness drop in adversarial training. *ArXiv*, abs/2010.08034.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Zhaohui Wy, Dawei He, Peng Cheng, Zhonghao Wang, and Haiying Deng. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *ArXiv*, abs/2311.15296.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and T. Zhang. 2023b. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *ArXiv*, abs/2309.06256.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. *ArXiv*, abs/2310.14566.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. *ArXiv*, abs/2303.02961.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2023c. Evalcrafter: Benchmarking and evaluating large video generation models. *ArXiv*, abs/2310.11440.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023d. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ArXiv*, abs/2310.05338.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. 2023. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *arXiv preprint arXiv:2312.11865*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *ArXiv*, abs/2308.09126.

Daniel McKee, Bing Shuai, Andrew G. Berneshawi, Manchen Wang, Davide Modolo, Svetlana Lazebnik, and Joseph Tighe. 2021. Multi-object tracking with hallucinated and unlabeled videos. *ArXiv*, abs/2108.08836.

Eric Melz. 2023. Enhancing llm intelligence with armrag: Auxiliary rationale memory for retrieval augmented generation. *ArXiv*, abs/2311.04177.

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

Carlos Mougan, Klaus Broelemann, David Masip, Gjergji Kasneci, Thanassis Thiropanis, and Steffen Staab. 2023. Explanation shift: How did the distribution shift impact the model? In *Arxiv*.

Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. 2022. Certified policy smoothing for cooperative multi-agent reinforcement learning. *ArXiv*, abs/2212.11746.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

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

Hao Peng, Hang Li, Lei Hou, Juanzi Li, and Chao Qiao. 2021. Multimodal entity tagging with multimodal knowledge base. *ArXiv*, abs/2201.00693.

Zeju Qiu, Weiyang Liu, Tim Z. Xiao, Zhen Liu, Umang Bhatt, Yucen Luo, Adrian Weller, and Bernhard Scholkopf. 2022. Iterative teaching by data hallucination. In *International Conference on Artificial Intelligence and Statistics*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

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 *Conference on Empirical Methods in Natural Language Processing*.

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. *arXiv preprint arXiv:2307.01928*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md. Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering. *ArXiv*, abs/2312.05200.

Bhaskarjit Sarmah, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali. 2023. Towards reducing hallucination in extracting information from financial reports using large language models. *ArXiv*, abs/2310.10760.

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900.

Anshul B. Shah, Aniket Basu Roy, Ketul Shah, Shlok Kumar Mishra, David W. Jacobs, Anoop Cherian, and Ramalingam Chellappa. 2023. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18846–18856.

Anton Shapkin, Denis Litvinov, and Timofey Bryksin. 2023. Entity-augmented code generation. In *Arxiv*.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *ArXiv*, abs/2108.13624.

Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. 2023a. Unified model for image, video, audio and language tasks. *arXiv preprint arXiv:2307.16184*.

Mustafa Shukor, Alexandre Ramé, Corentin Dancette, and Matthieu Cord. 2023b. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *ArXiv*, abs/2310.00647.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825.

Arvind Krishna Sridhar, Yinyi Guo, Erik Visser, and Rehana Mahfuz. 2023. Parameter efficient audio captioning with faithful guidance using audio-text shared latent representation. *arXiv preprint arXiv:2309.03340*.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023a. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs? *ArXiv*, abs/2308.10168.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023b. Aligning large multimodal models with factually augmented rlhf. *ArXiv*, abs/2309.14525.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. 2023a. Large language models as generalizable policies for embodied tasks. *ArXiv*, abs/2310.17722.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. 2023b. Large language models as generalizable policies for embodied tasks. *arXiv preprint arXiv:2310.17722*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023c. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Nasib Ullah and Partha Pratim Mohanta. 2022. Thinking hallucination for video captioning. In *Asian Conference on Computer Vision*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Andrés Villa, Juan Carlos Le'on Alc'azar, Alvaro Soto, and Bernard Ghanem. 2023. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *ArXiv*, abs/2312.02219.

Junyan Wang, Yi Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Mingshi Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023a. Evaluation and analysis of hallucination in large vision-language models. *ArXiv*, abs/2308.15126.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *ArXiv*, abs/2311.07397.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023c. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *ArXiv*, abs/2312.01701.

Jing Wu, Jennifer Hobbs, and Naira Hovakimyan. 2023. Hallucination improves the performance of unsupervised visual representation learning. *ArXiv*, abs/2307.12168.

Wenke Xia, Dong Wang, Xincheng Pang, Zhigang Wang, Bin Zhao, and Di Hu. 2023. Kinematic-aware prompting for generalizable articulated object manipulation with llms. *arXiv preprint arXiv:2311.02847*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Jiao Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *ArXiv*, abs/2306.09265.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. 2023. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Munan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *ArXiv*, abs/2310.01469.

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

Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Changdong Yoo. 2022. Information-theoretic text hallucination reduction for video-grounded dialogue. In *Conference on Empirical Methods in Natural Language Processing*.

Fan Yu, Haoxu Wang, Ziyang Ma, and Shiliang Zhang. 2023a. Hourglass-avsr: Down-up sampling-based computational efficiency model for audio-visual speech recognition. *arXiv preprint arXiv:2312.08850*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chun yan Li, Zheyu Zhang, Yushi Bai, Yan-Tie Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yun Peng Qi, Weikai Li, Yong Guan, Kaisheng

Zeng, Ji Qi, Hailong Jin, Jinxi Liu, Yuxian Gu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023b. Kola: Carefully benchmarking world knowledge of large language models. *ArXiv*, abs/2306.09296.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490.

Yongsheng Yu, Heng Fan, and Libo Zhang. 2023d. Deficiency-aware masked transformer for video inpainting. *arXiv preprint arXiv:2307.08629*.

Zhongjie Yu, Shuyang Wang, Lin Chen, and Zhongwei Cheng. 2023e. Halluaudio: Hallucinate frequency as concepts for few-shot audio classification. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *ArXiv*, abs/2106.11520.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Annual Meeting of the Association for Computational Linguistics*.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023a. Halle-switch: Controlling object hallucination in large vision language models. In *Arxiv*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Y. Ma. 2023b. Investigating the catastrophic forgetting in multimodal large language models. *ArXiv*, abs/2309.10313.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

Chen Zhang. 2023. User-controlled knowledge fusion in large language models: Balancing creativity and hallucination. *ArXiv*, abs/2307.16139.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

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

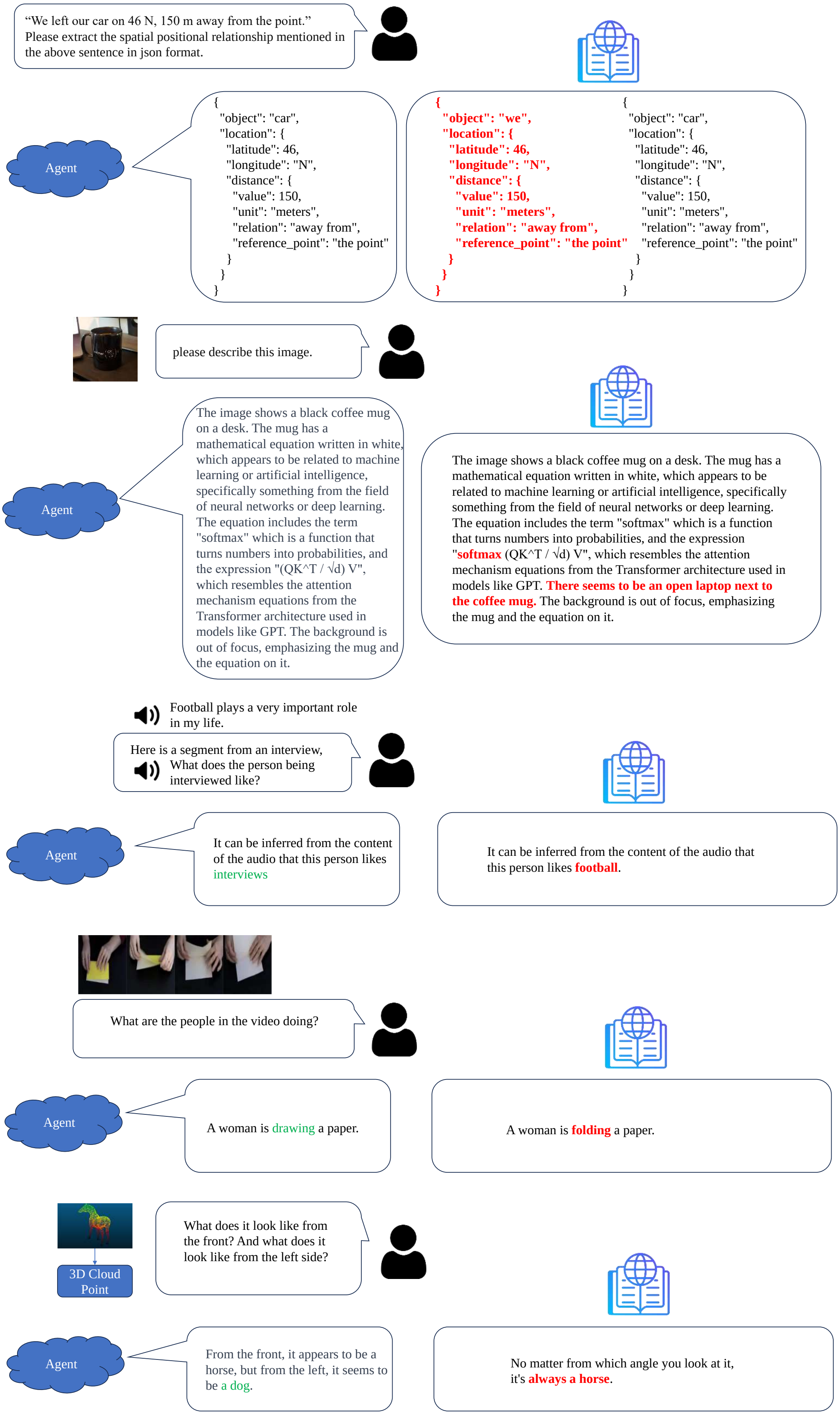
Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. 2021. Boosting entity-aware image captioning with multi-modal knowledge graph. *ArXiv*, abs/2107.11970.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiao wen Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *ArXiv*, abs/2311.16839.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023b. Analyzing and mitigating object hallucination in large vision-language models. *ArXiv*, abs/2310.00754.

A Example Appendix



"We left our car on 46 N, 150 m away from the point."
Please extract the spatial positional relationship mentioned in the above sentence in json format.

```
{
  "object": "car",
  "location": {
    "latitude": 46,
    "longitude": "N",
    "distance": {
      "value": 150,
      "unit": "meters",
      "relation": "away from",
      "reference_point": "the point"
    }
  }
}
```

```
{
  "object": "we",
  "location": {
    "latitude": 46,
    "longitude": "N",
    "distance": {
      "value": 150,
      "unit": "meters",
      "relation": "away from",
      "reference_point": "the point"
    }
  }
}
```

```
{
  "object": "car",
  "location": {
    "latitude": 46,
    "longitude": "N",
    "distance": {
      "value": 150,
      "unit": "meters",
      "relation": "away from",
      "reference_point": "the point"
    }
  }
}
```

please describe this image.

The image shows a black coffee mug on a desk. The mug has a mathematical equation written in white, which appears to be related to machine learning or artificial intelligence, specifically something from the field of neural networks or deep learning. The equation includes the term "softmax" which is a function that turns numbers into probabilities, and the expression "(QK^T / √d) V", which resembles the attention mechanism equations from the Transformer architecture used in models like GPT. The background is out of focus, emphasizing the mug and the equation on it.

The image shows a black coffee mug on a desk. The mug has a mathematical equation written in white, which appears to be related to machine learning or artificial intelligence, specifically something from the field of neural networks or deep learning. The equation includes the term "softmax" which is a function that turns numbers into probabilities, and the expression "**softmax** (QK^T / √d) V", which resembles the attention mechanism equations from the Transformer architecture used in models like GPT. **There seems to be an open laptop next to the coffee mug.** The background is out of focus, emphasizing the mug and the equation on it.

🔊 Football plays a very important role in my life.

Here is a segment from an interview, 🔊 What does the person being interviewed like?

It can be inferred from the content of the audio that this person likes interviews

It can be inferred from the content of the audio that this person likes **football**.

What are the people in the video doing?

A woman is drawing a paper.

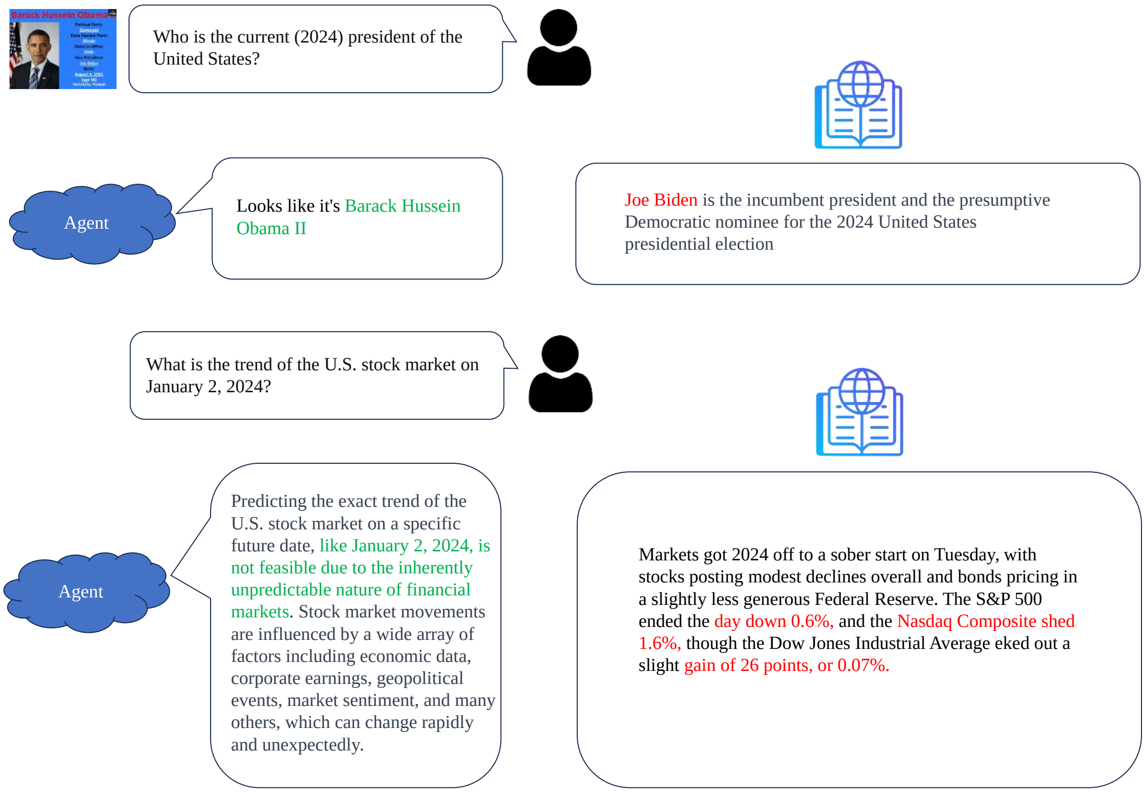
A woman is **folding** a paper.

What does it look like from the front? And what does it look like from the left side?

From the front, it appears to be a horse, but from the left, it seems to be a dog.

No matter from which angle you look at it, it's **always a horse**.

Figure 4: Conflict in Intrinsic Knowledge examples.



Who is the current (2024) president of the United States?

Looks like it's Barack Hussein Obama II

Joe Biden is the incumbent president and the presumptive Democratic nominee for the 2024 United States presidential election

What is the trend of the U.S. stock market on January 2, 2024?

Predicting the exact trend of the U.S. stock market on a specific future date, like January 2, 2024, is not feasible due to the inherently unpredictable nature of financial markets. Stock market movements are influenced by a wide array of factors including economic data, corporate earnings, geopolitical events, market sentiment, and many others, which can change rapidly and unexpectedly.

Markets got 2024 off to a sober start on Tuesday, with stocks posting modest declines overall and bonds pricing in a slightly less generous Federal Reserve. The S&P 500 ended the day down 0.6%, and the Nasdaq Composite shed 1.6%, though the Dow Jones Industrial Average eked out a slight gain of 26 points, or 0.07%.

Figure 5: Factual Conflict examples.