Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multi-agent Court to Mitigate VLM Hallucinations

Percy Lam

ph125@cam.ac.uk

University of Cambridge

Lavindra de Silva

University of Cambridge

Weiwei Chen

University College London

Ioannis Brilakis

University of Cambridge

Research Article

Keywords: vision language model, multi-agents, hallucination, argumentation

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8949950/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multi-agent Court to Mitigate VLM Hallucinations

Percy Lam[1*], Lavindra de Silva[1], Weiwei Chen[1,2], Ioannis Brilakis[1]

[1*]Department of Engineering, University of Cambridge, 7a JJ Thomson Avenue, Cambridge, CB3 0FA, United Kingdom.
[2]Bartlett School of Sustainable Construction, University College London, 1-19 Torrington Place, London, WC1E 7HB, United Kingdom.

*Corresponding author(s). E-mail(s): phl25@cam.ac.uk;

Abstract

Hallucinations have hindered the widespread use of vision language models (VLMs) for domain-specific applications such as road maintenance. While previous researchers constructed multiple solutions for different sources of visual hallucinations, knowledge gaps persist in handling context-dependent hallucinations where the targeted objects are difficult to be prompted precisely. This research explores hallucination by converting image-to-text binary classifications into evidential arguments by VLM agents, each providing a binary Yes/No answer with a justification. The proposed solution involves VLM agents performing distinct roles, starting with a detection unit that uses a primary detector and a reviewer to verify scope compatibility. These agents interact to aggregate their findings and justifications into a single, unified verdict. The different roles of each agent are inspired by the distinctive roles of the prosecutor, the defence counsel and the judge, while the questioning techniques used by the justification reviewer are inspired by lawyers' examination techniques in court room and argumentation schemes. Experiments are performed on toppled poles from road scene images in the Urban Issue dataset, and wider general adoption through subsets of the PhD dataset annotated on COCO2014. Experiments show that our solution achieved a superior overturning rate of 30% and 2.3 percentage points increase in F1 score in the domain-specific application, with 50% less time required than the closest multi-agent solution. Comparable detection performance and efficient resource consumption were also seen in the general adoption.

Keywords: vision language model, multi-agents, hallucination, argumentation

1 Introduction

Vision language models (VLM) have seen dramatic advances in the past several years. The ability to use VLM to perform Vision Question Answering (VQA) attracts interest for many potential domain-specific applications, such as smart manufacturing [1], medical reporting [2] and road defect detections [3]. Incorrect or unfounded responses outputted from VLMs, collectively known as hallucinations [4], have been a barrier to adopting VLM in daily business operations. The technology industry has implemented various methods to mitigate hallucinations, such as verifying with external web information [5], chain of thought reasoning [6] and using external solvers [7], which have improved the grounding and authenticity of outputs. Despite ongoing research and enhancements, hallucinations remain an impediment to more practical applications as users seek more explainability from VLM. Combining VLM with established concepts from multi-agent systems research, particularly argumentation, is emerging as a promising direction for tackling such challenges.

This research explores performing image classification tasks through multi-agent argumentation. To support this, we require VLM to output a conclusion statement supported by a justification as a premise, as opposed to a simple binary Yes/No output without an explanation, which users/agents cannot easily inspect and correct. By having a conclusion and a premise, the VLM agents are allowed to launch argumentative attacks to combat hallucinations, such as directly rebutting the conclusion, undermining the premise and undercutting the connection between the two. Argumentation provides a standardised analytical framework to reconcile diverse agent outputs, ensuring the process of challenging and defending evidence leads to more robust and accurate detections.

Since hallucinations in VQA often arise from the perception of VLM agents and the inference from the perceived facts to the conclusion, the arguments presented by the VLM agents are similar to evidential arguments [8] in law for judging evidence and witnesses. This branch of argument differs from rule-based or case-based reasoning [9, 10]. Evidential arguments are used to help VLM agents interpret conflicting, incomplete or uncertain evidence. This includes determining the attributes or conditions of the targeted object by using different agents to detect or check the relevant regions in an image.

Rather than strict binary true or false conclusions, real-world object conditions often involve inherent ambiguity and context-specific nuances. The visual element introduces a layer of complexity and uncertainty beyond traditional text-based reasoning, as the model's cross-modal attention shifts across the image in response to specific text prompts. For instance, VLM subjected to images of toppled poles on highways frequently yielded ambiguous results, such as claims of poles "leaning but not fully toppled" or confusion between wooden poles and trees. The visual and textual inputs resulting from an evidential argument allow the VLM to qualitatively critique textual justifications and a freedom to attack arguments at various strengths, including to overrule an argument or to undermine a conclusion for an aggregator to decide. Presenting detection as an argument also allows flexibility in presuming the detection to be true until shown otherwise, analogous to defeasible arguments [11, 12], as opposed to a strict binary conclusion.

Because of the way that classifications become evidential arguments, agents are also required to take different roles and interact in a certain way. Especially with how VLM works, assigning the VLM too many tasks within the same query prompt is prone to receiving hallucinated answers or missing different parts of the tasks [13]. Functional specialisation and normative specification would be necessary for agents to perform certain roles within the bounds of certain authority and constraints in communication [14]. The rules in assigning roles, matching and enacting agents with roles in a potentially open agent system were discussed, for example in [15]. In agentic AI, LLM and VLM agents commonly take part in various roles, collaborate in a pre-programmed manner or are planned by LLM in-context or post-training, and seek improvement through test-time interactions [16].

This research approaches contextual hallucination in VLM detections through the lens of building a road defect classifier to identify toppled poles from road scenes. Toppled poles are one of the highest-risk road incidents that require response within one hour of reporting [17], making the ability to detect them a key step towards automating road maintenance. Beyond road maintenance, the proposed solution can also be applied to more general problems, as verified by further experiments with hallucination evaluation datasets (subsets of the PhD Dataset [18]) annotated on COCO2014 [19]. This research demonstrates an avenue to improve VLM hallucination mitigation by applying concepts of argumentation and roles in a multi-agent system, and how these benefit a real-world problem such as road maintenance. In terms of contributions, this paper:

- Translates a binary image classification problem into an evidential argument to allow human users to inspect the decision-making and mitigate hallucinations by launching attacks when building VLM multi-agents
- Reduces contextual hallucination where an answer may be right in general but not compatible with the desired scopes or attributes (such as defining electrical, utility and traffic poles in finding "toppled poles"), and is difficult to resolve by direct prompting. This is achieved by treating detection results as defeasible arguments to be examined by a novel pipeline of VLM agents in different roles, separating detection and in-scope justification review. This can be achieved in domain-specific applications, such as finding toppled poles, and general applications of verifying attributes of generic objects.
- Provides a novel, intuitive and minimal pipeline that achieves superior detection performance, time taken and cost than other multi-agent setups as shown through experiments in both domain-specific (Urban Issue) and general (PhD attribute) datasets. The pipeline is also model agnostic, training-free and does not require an external object detection or foundation model to cue the prompt for VLMs.
- Preserves ambiguity and nuance in the justification and prevents a lost of contextual details when reviewing and aggregating responses by using qualitative attributes written in natural language texts.

2 Related Works

Research on hallucinations has been rapidly growing in both VLMs and large language models (LLM). In VLMs, the primary focus has been on reducing hallucinations for general VQA and object detection. In LLMs, researchers have explored hallucination in a broader range of applications, such as code generation [20, 21] and multi-agent interaction [22–24]. In more traditional artificial intelligence, argumentation provides an analytical framework to understand arguments, commonly applied in specialist domains such as law.

Amongst past studies in VLMs, researchers identified several major causes of hallucinations. **Modality misalignment** between the vision and language components of the query was frequently highlighted. For example, the vision component may be suppressed [25, 26] or over-rely on the textual component when making decisions [27–29]. During inferences, the models could suffer from **knowledge and context inadequacy** that they often relied on pre-existing knowledge [30, 31] and failed to interpret and resolve conflicts with the provided context [32, 33]. This aggravated other deficiencies, such as fading memory [30, 34], poor perception of the input image [35] and suboptimal linkages between the perception and internal knowledge [33].

Apart from the inherent handicaps in VLMs, responses may be hallucinated by **deficiencies of individual models**. The models may fail to differentiate hallucination [36–38] and provide ungrounded generation [39, 40] or unfaithful outputs [41–43]. In LLM, research pointed to a lack of verification [44, 45] and inconsistencies in answers [46–48]. Sometimes unsatisfactory answers were also caused by a lack of cues to reach answers [49] and making insufficient [35] or implausible [50] reasoning steps. A lack of explainability [51] and quantification of helpfulness and truthfulness [52] also constrain hallucination detections.

2.1 Literature on resolving hallucinations

Ungrounded outputs are traditionally resolved by finding references with retrieval techniques [21, 22] and improving the inference model. Generic models improve for specific domains by knowledge distillation [53, 54], supervised fine-tuning [32, 36, 43] and providing extra guidance during training and inference [23, 49, 55, 56], potentially through verifying with external solvers over ground truths [7]. Some attempted to address model deficiencies by examining different parts of the model [38, 57] and subsequently improving the encoder [39] and/or the decoder by contrast decoding [27–29, 31, 37]. The outputs can also be aligned by probabilities, divergence or losses [30, 33, 42, 47].

Multi-agentic AI were increasingly gaining traction in resolving hallucinations. Multiple LLM agents could be employed to perform iterative reviews [40, 58, 59], debate to seek consensus [24, 35, 44–46, 50] or both [25]. Employing multiple agents also enabled division of tasks [20] or data segments [36].

2.2 Literature on applying argumentation

While arguments are commonplace in daily lives, argumentation in computer science provides a structured way to study arguments [9]. It typically comprises a conclusion,

a set of premises, and an inference rule from the premises to the conclusion [11]. It can be used to persuade or critique proposals in negotiations, or as an independent mechanism to understand reasoning by agents [60]. Arguments can even be drawn in argument diagrams to help understand the underlying logic better and identify argumentation schemes for commonly used types of arguments. As opposed to monotonic reasoning or deductions where a conclusion is conclusively valid, there are defeasible arguments which are fallible upon new, stronger arguments or premises. While a defeasible argument can be accepted tentatively with the currently available evidence, this may be retracted when new evidence appears [61].

Various frameworks of formal argumentations were proposed on the abstract level between arguments. Abstract argumentation is pairs of arguments and attack relationships between arguments proposed by Dung [62]. Bipolar argumentation framework extends on Dung to allow both attack and support. Tripolar argumentation also allows neutralisation [63]. Quantitative bipolar argumentation framework evaluates arguments with a gradual evaluation with quantified intrinsic strengths [64].

These argumentation frameworks were applied in various proposed solutions to generate controlled model outputs, transparent predictions and allow structured modifications upon human interaction. For example, on the topic of argument mining, it was shown possible to extract argumentative elements and construct argument graphs from non-argumentative sources such as movie reviews on Rotten Tomato [65], and textual health reports where transformer-based models were used to assist clinicians in decision making [66]; other application areas include misinformation detection and scientific literature analysis [67]. Hybrid approaches also exist that combine formal argumentation theory with data-centric approaches for argument extraction, for example to estimate the helpfulness of Amazon product reviews [68], predict the winning stance in professional natural language debates [69], and provide predictions on both images (faces and food items) and text (Amazon and IMDB reviews), with explanations used to underpin predictions [70].

On the topic of argument generation, there are rule-based techniques that combine mined argument elements, represented as logical recipes or graphs, to create complete arguments using approaches such as hierarchical planning [71], and Bayesian propagation to update the agent's confidence in the premise and its supporting beliefs as arguments are progressively generated [72]. There is also work on using argumentation schemes when specifying norms as part a multi-agent system design [14] to enhance control on agent communication and constrain agent moves and norms violations, as well as research on making argument generation conform to normative requirements or a desired user strategy (e.g. the need for the shortest argument or those that are the most explainable), as surveyed in [67].

To eliminate the need to curate or search datasets when generating arguments, recent work has increasingly adopted LLMs, seeking a trade-off between reliability and the novelty and naturalness of generated arguments (see, for example [73]). Such uses of vision/language models are rapidly becoming more sophisticated, extending to multi-LLM and multi-agent settings, such as fine-tuning a VLM on domain-specific data to enable assessing and explaining claims arising from chart interpretation [74]. Further on reviewing the validity of claims, research sought to uphold response consistency

by enabling LLMs to self-check their own claims [40, 47], or developing debating and voting agents to reduce hallucinations in generated claims [45, 46].

Instead of resolving conflicts between arguments, there exists an intermediate level to examine individual elements and connections of arguments [12]. Evidence-based reasoning is often applied to interpret conflicting, incomplete or uncertain evidence [8], such as courts to settle questions of facts from expert opinion or witness testimonies [10]. By making an argument with supporting premises, various argumentation schemes [11, 75] are available in attacking various weaknesses of the premise (evidence) or the inference from the premise to the conclusion. Attacks towards different elements of arguments were demonstrated in a worked example of speeding with various exceptional circumstances [76].

Reviewing the gaps in knowledge, there is plenty of research on using multiple thinking steps and/or multiple agents to find corroboration of answers to eliminate hallucinations. This research problem is different because the answer is not strictly right or wrong as in Sun [45], but is judged on the context and explanation. The answer is not sufficient to simply be consistent across repeated predictions [47] or voted by a simple majority [46]. Even if a VLM detects an object and its attributes, then uses the attributes to query for the object and form a closed loop as Wu [40], an answer can still be wrong if the justification is wrong. Evaluating justifications outputted by VLM/LLM remains scarcely adopted in finding hallucinations.

As opposed to conflicts between arguments, the problem of identifying toppled poles or other attributes in the generalisation test concerns the evidence and its inference to a conclusion. The query for toppled poles is therefore not easily decomposed into sub-problems [44] or summarised in a score [22, 58] to be solved in a divide-and-conquer approach. By adding more qualifiers or more complex concepts into the prompt, VLMs will become less accurate in VQA or vision reasoning [77] and more prone to getting falsely aligned answers. The research therefore needs to explore an approach to start with a straightforward query, supported by justification evaluation by multiple agents, to arrive at an explainable answer.

Amidst a plethora of research on assembling VLM agents and inputting guidance objects in prompts to perform specific tasks or roles, there is a need to apply an analytical framework to underpin the rationale in constructing agentic AI architecture. Previous research applying (abstract) argumentation focused on extracting features or sub-features of arguments and forming decision-making rules. While they (especially from non-argumentative sources) were necessary before the prevalence of LLM and VLM and remain influential in providing controllable decision flows, the advent of the LLM and VLM encourages more exploration in making use of the generated conclusion and premise to polish performance downstream.

3 Methodology

The problem statement in Section 1 illustrates the research goal of identifying toppled poles by finding detection results with justifications compatible with its more detailed description. VLMs were found to create some discrepancies between the binary Yes/No conclusion and the explanation, caused by confusion between wooden poles and trees

or omissions of objects. The discrepancies inspire the proposed solution to designate separate VLM agents to perform detection and review. Figure 2 illustrates the proposed solution, which includes a detector, a reviewer and an aggregator. Each module is a separate chatbot calling and receiving from a VLM API, being assigned the role through natural language text prompts and receiving outputs from the upstream module. The setup is inspired by the judicial system, where a case is first advanced by the prosecution or the plaintiff and challenged by the defence lawyers on errors of fact or inference. The judge considers arguments from both sides and makes the final verdict with an explanation. The concept of courtroom litigation is translated into a unidirectional argument across multiple VLM agents performing different roles.

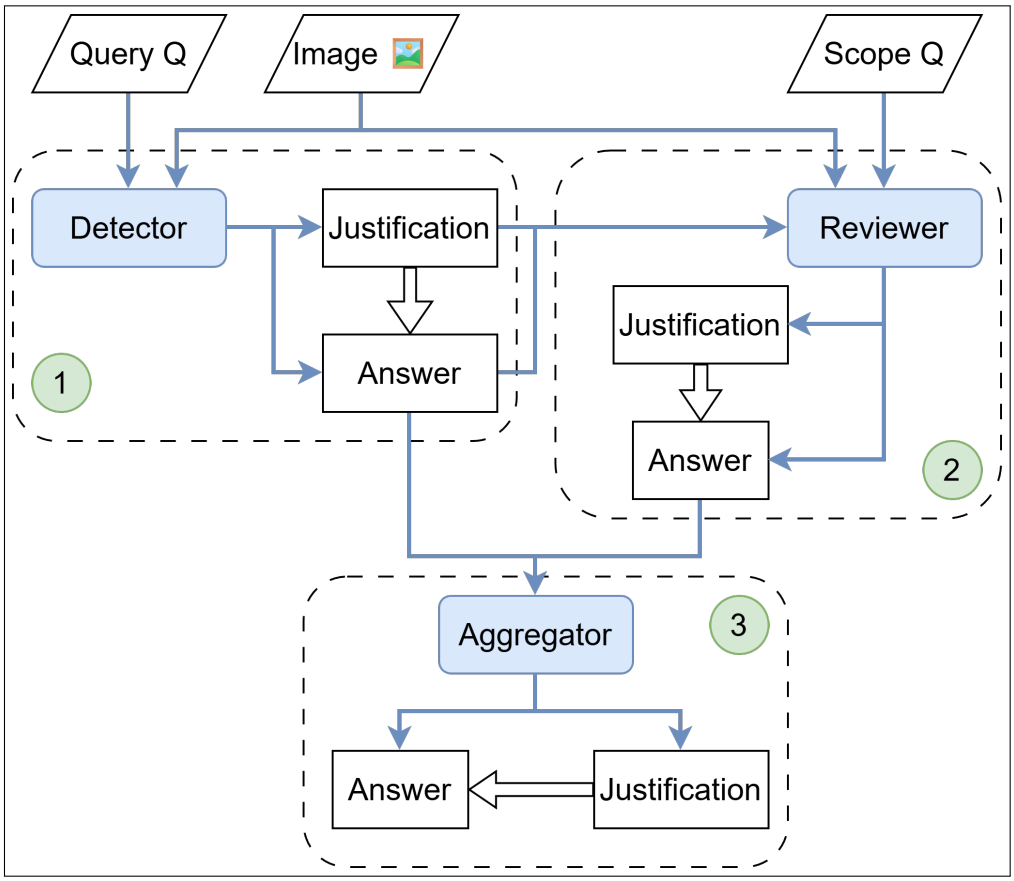


Fig. 1 The proposed solution that (1) inputs the query image and question into a detector, producing an initial answer supported by a justification; (2) feeds the answer and justification to the reviewer, together with a scope with more well-defined attributes of the targeted object, for an answer and justification; (3) provides both sets of answer and justification to the aggregator for the final decision, supported by a summarised justification

3.1 The detector and the reviewer

The detector is the first point of call to query about the image and receive a binary Yes/No answer and a justification. These are brought to **the reviewer** to inspect whether the detected object falls within a scope. The scope captures the user-defined object attributes for domain-specific applications, or for general applications, reviews whether the detector provides justifications that are compatible with the reviewer's own explanation on the same image. At the end of the detector and the reviewer, **the aggregator** receives a dictionary containing the image name, the binary Yes/No outputs and the justifications of both modules. The prompts used in each module are recorded in the Section A.

The reviewer is the main component to eliminate contextual hallucinations of out-of-scope answers. The module receives the binary answer and the justifications from the detector and reviews whether the detected object attribute aligns with the defined scope. The scope covers toppled electrical, utility and traffic poles, while excluding trees and unrelated pole defects. To further constrain the VLM chatbot to review the argument instead of performing a fresh detection, the text prompt explicitly reminds it to only use the image as a context for the textual justification and not to repeat detection. When classifying for general applications, without knowing the objects referenced in the query question in advance, the reviewer first detects objects in the image. It then cross-checks whether the detector (tasked to find the presence or absence of a particular object or attribute in the query question) provides an answer compatible with its own and relevant to the query question, and outputs the reasoning for the compatibility check.

Unlike traditional pre-programmed symbolic agents, VLMs can suffer from attentional drift on long prompts. **The division of roles** in each module avoids a VLM from performing too many tasks at once and losing attention to the particulars. The role division enables more fine-grained control on each module to best address the perceptual and contextual deficiencies and facilitates end-users to find errors more easily. For instance, the reviewer will have the benefit of having more descriptions of the desired object and self-query whether the object is present or not, with questioning styles based on the detector's answers. The benefit of dividing roles through text prompting is that the same foundation VLM can be used for the detector and the reviewer, leaving substantial flexibility to adapt across different tasks without the need to reconstruct the agents.

The presence of arguments allows different **questioning styles and attacks** under different argumentation schemes. The detector's answer is essentially a defeasible argument ready to be attacked and potentially defeated by the reviewer. When the reviewer receives a 'Yes', the reviewer is tasked to attack the premise, the inference rule and the conclusion, analogous to cross-examinations by lawyers of the opposite side in a courtroom. When the answer is a 'No', the reviewer assists the detection with open questions to find potentially omitted objects, analogous to lawyers of the same side asking questions to help witnesses recall memories. This has to be coupled with a prohibition not to force a 'Yes' answer, as the recollection questions are merely assistive in case of absent-mindedness. There are several other established attacking techniques, such as the critical questions for arguments from circumstantial evidence

or witness testimony [75]. These critical questions allow attacks to be focused on the points of contention, not airing attacks on everything that simply wastes time [10]. A trial was performed with further discussions in Section 5.4.

Multiple other argumentation configurations have been considered. These configurations will be discussed in the ablation studies (Section 5.2) to examine the extent of improvement brought by justifications set up in different ways.

- Answering with a generic warmup question: caters for another feature of VLM, where providing contexts helps answer a question [56]
- Self-query: instead of calling separate VLM agents, the same agent will be injected the scope and ask to reflect itself, having previous answers remaining as contexts. This extrapolates on the VLM feature on contexts, as well as examining whether corroboration of previous arguments exists.
- Adversarial: two independent detectors make their own arguments and present them to the aggregator, effectively acting as a judge. In a more extreme variant, the two independent detectors are instructed to provide opposing conclusions with their supporting premises.

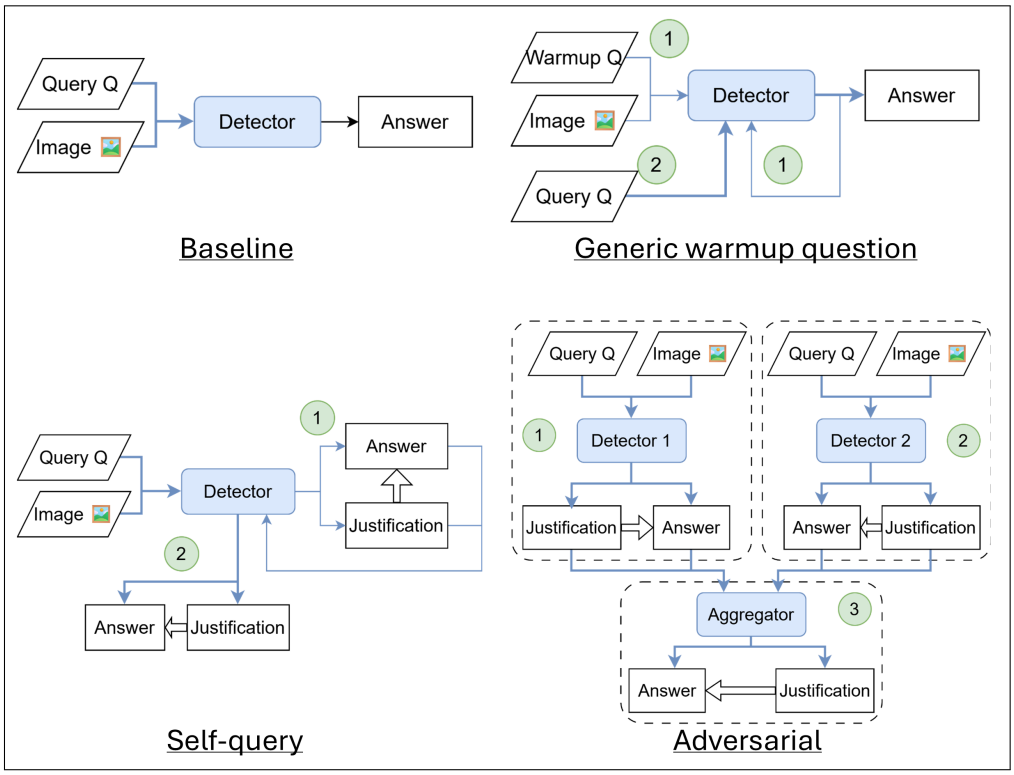


Fig. 2 The four ablation setups: (1) baseline with plain query question; (2) answering with a generic warmup question; (3) self-query; (4) adversarial debater setup

Other agentic AI pipelines proposed by others will be compared in the comparative study in Section 4.1.

3.2 The aggregator

The aggregator receives a list of JSON dictionaries from the detector and the reviewer for the same image. It then summarises the responses to give a final binary Yes/No answer, the dissenting module and the justifications for outputting the binary answer. This is an important part of the multi-agent approach because it reads the different viewpoints and independent reasoning of each module, uses a partially constrained rule to make a decision and exports the final verdict.

The aggregator performs its role by receiving a ground rule specifying how it uses information from other agents and its final objective. The rule is slightly different in the comparative study with the domain-specific dataset than in the general dataset because having the target attributes of the toppled poles allows the aggregator to evaluate not only the compatibility between the detector and the reviewer, but also the compatibility with the defined attributes in the scope. This is mitigated by re-injecting the query question into the aggregator, so it has an idea of the attributes of what object to consider.

4 Experimental setup

Experiments were performed to compare against the baseline setup that asked the plain question [1]. Apart from the detector-reviewer-aggregator setup as illustrated in Section 3, several multi-agentic AI setups were compared. The comparative setups were compared on the domain-specific dataset of toppled poles and a general dataset on attributes of various objects.

Experiments on the comparative and ablation studies were run on Gemini-2.5-flash-lite [78]. Our preliminary studies Section B showed that not only did the model show superior latency and cost-effectiveness to the more complicated flash or pro model, but it also returned better detection results in the baseline. This may potentially be caused by less self-confusion during chain of thought thinking.

4.1 Setups for the comparative study

The comparative study comprised several setups. These setups highlighted modes of communication by VLM agents and how they impacted the effect of mitigating hallucination. The experimented setups included:

- Anchoring and feedback (Kwartler Mod): Anchoring is a priming behaviour in human brains. It happens because System 1 of a human brain likes associating objects with like memory, evoking selective memory. [79] This setup was previously proposed by Kwartler [59] on LLMs, which aimed to reveal how having the first detection and feeding back the second reviewed detection would influence the final argument by the detector. This differed from our proposed setup, which used a third agent without prior detections to make a final decision.

[1] Does the image contain any toppled electrical poles? Please answer with a justification.

- Debates (Du Mod): Both agents would first make independent detections. The debater received the scope so its response would have considered the more detailed descriptions of the desired object. The agents then exchanged their responses and continued to debate until they reached the same Yes/No answer (or a maximum of 10 iterations for computational constraints). When a consented Yes/No answer was reached, an aggregator would summarise an overall justification without affecting the Yes/No answer. This was inspired by Du [50]
- Debates with guidance (Lin mod): Adapted from Lin [35], instead of having unconstrained and unguided debates, after both agents embarked on their independent detections, the agent that responded 'Yes' was inquired further about the detected object and its attributes (shapes). The other agent would try to find the same object and feed the reasoning back to the first agent for confirmation. In case the two agents disagreed on the detected object, a third agent (a judge) made the final decision.
- Self-consistency, history across images and voting between agents (Yang mod): adopted from Yang [46], each deployed agent would first detect the same image several (set as 5) times. The output Yes/No answer and the justification would be parsed into an encoder (gemini-embedding-001) to retrieve an embedding to find outliers deviating from the mean. The summarised answer across all self-enquired rounds per agent would be presented for voting. The voting weight of the dissenting agent would diminish until a simple majority decision was reached. The dissenting agent would be reminded in the next image that it diverged from the consensus and cautioned to align with other agents.

The comparative setups are summarised in Table 1. The temperature of the VLMs was set to 0, except for setups with debaters (Du mod and Lin mod), where the debater was set to 1 to provide some variation in answers.

# Agents	Mechanism	Measured Result	Prior Work
2	Effects of anchoring and feedback: A predicts, B justifies, and A refines the answer	A's refined answer	Kwartler
2	Effects of debates for a negotiated answer: - A and B both predict. Exchange predictions and negotiate until both agree	The final answer	Du
3	Effects of debates with guidance: - A and B both predict. The agent that returned a positive detection (say A) was asked about more attributes (shapes) of the predicted object - The other agent (B) checks if it finds the predicted object and if so, why - The agent that returned a positive detection (A) counterchecked • if it can find its previously predicted object • if it thinks it's A's or B's predicted object and why - A final judge C makes the final decision and reason	C's answer	Lin
3	Effects of (1) history across images, (2) self-consistency and (3) voting - Repeated self-query and flag outlier responses as 'hallucinations' - Vote to find a consented answer and reduce the weight of the dissenting agents in every voting round - Add prompt to make the wrong agent reflect and avoid mistakes Note: History reset in batches of 10 images to stay within the VLM's context window. Batchwise optimisation with entropy was not pursued	The final answer	Yang

Table 1 Adaptations of prior work on multi-agent debates for the comparative study. The scope for justification reviews was injected at different agents of the setups.

	Ground Truth	
Toppled pole?	Yes	No
pole	385	128
tree	12	970

Table 2 The ground truth distribution of the experimented subset of the Urban Issue Dataset

4.2 Datasets

Experiments were performed on two datasets. The solution was developed, compared with past studies and benchmarked on multiple VLMs on the domain-specific Urban Issue Dataset [80]. Additional comparative experiments with past studies were performed on a subset of the PhD dataset [18], which was prepared on COCO2014 [19], to demonstrate the solution's efficacy for general applications.

The domain-specific evaluation was conducted on relevant parts of the Urban Issue Dataset. The full dataset was shared on Kaggle, covering 10 categories of common urban road defects and hazards. Experiments were conducted on the validation sets of the "Fallen trees" and "Damaged Electric wires and poles" categories with 1495 images in total. The "Damaged Electric wires and poles" category contained not only toppled poles but also other defects, such as bird nests on top of the utility poles, obstructions hung on electric cables and smoking transformers. A score of images in "Fallen trees" concurrently showed toppled poles in the same image frame. The two validation sets were manually reviewed and summarised with the ground truth distributions in Table 2.

The evaluation for general applications was performed on a subset of 2000 random images for object attributes in *PhD-base*. This subset of the PhD dataset was chosen because it contained query questions on the attributes of objects in the ground truth, which was more sophisticated than the presence of objects presented in the original COCO2014 dataset. Sometimes the query questions asked about objects in the periphery instead of the main object, increasing the difficulty of the task. The dataset was annotated with both 'yes_question' and 'no_question', expecting 'Yes' and 'No' answers respectively. Both were reviewed to remove ambiguous answers and were tested with results in Section 5.3. The challenge in ascertaining the attributes of objects in PhD-base was analogous to differentiating kinds of toppled poles.

4.3 Metrics

Results were gathered from the final Yes/No answer of each experimental setup in the comparative and the ablation study. Each setup was run 5 times to seek an average precision, recall, accuracy and F1 score. These formed the basis of comparing the detection performance.

The filter rate metric was used as part of the evaluation to demonstrate the effect of overturning initial wrong detections with the reviewer and the aggregator.

$$\text{Filter Rate} = \frac{\text{Filter Success}}{\text{False Detection}} \tag{1}$$

False detection counted images where the detector flagged an incorrect answer. **Filter success** counted images where the detector flagged an incorrect answer, but the final answer was reversed and complied with the ground truth.

Besides detection performance, the experiments also compared the efficiency. The time taken per image and the cost (£) per trial were compared. A usable architecture for practical applications shall provide good detection at a reasonable cost and speed. The latency per API call was provided for reference to show the contribution of time cost by the model, as opposed to the algorithm or data in/out in the computer performing the tests.

5 Results

Our proposed solution outperformed the baseline, comparative and other ablation setups in the detection performance with superior time and cost. Detailed detection results for comparative and ablation studies were shown in Table 3 and Table 4. Our proposed solution provided impactful overturns from the initial detection, where out of 24 overturn attempts, 21 correctly reversed false detections. Overturns were more effective in overturning false positives into true negatives. This was predominantly seen in falsely detecting toppled trees as poles. As revealed by the communication log, the detector often saw the toppled object as a wooden pole. By injecting the justification explaining what was seen, the reviewer paid attention to the patches of the toppled object and argued that the texture and appearance of the toppled object would reasonably infer it to be a tree truck and not a wooden pole. The aggregator could pick up the nuances and the details of the descriptions, and rightly overturn the false detection. On the other hand, the bulk of the unaddressed false detections were mainly false negative poles and false positive trees. This may have indicated perceptual shortfalls of the VLM in misinterpreting the object or failure to see the object.

5.1 Comparative study with the domain-specific dataset

Results in Table 3 showed our proposed solution outperformed other setups in the detection performance with superior time and cost.

Compared with **Kwartler mod**, the metrics and overturning patterns were similar to our proposed solution, with both precision and recall lower than ours by approximately 1 percentage point. Out of the handful of different determinations (28 in 1495 images), this mechanism with feedback made more inconsistent detections with poles (5 false positives and 6 false negatives) and more false positives with trees, despite other successful overturns.

In debating setups of **Du mod** and **Lin mod**, the recall was higher than the proposed solution but the precision dropped off significantly. The debating setups were found to have a tendency to generate positive outputs, a common hallucination also reported by Chu [22] when multiple VLMs collaborated. Among the images with divergent responses between agents (110 nos. in Du mod), debates did not appear to aid in finding out-of-scope hallucinations. Instead in most of the divergent responses, the detector conceded to align with the debater and returned a 'Yes' answer in the second iteration, even if their initial 'No' response might be correct. For images that required

multi-round debates, such communications appeared not to advance reasoning but instead reinforced each other on a hallucinated finding, swapped findings between two agents, or devolved into a deadlock of rhetorical standoff.

The communication log from Lin mod showed guided debates led to more useful debates than a free debate in Du mod, which partly explained a slightly better detection performance. In the communication logs, when both detectors disagreed with the finding, explicitly extracting the detected object helped the dissenting agent to narrow down perception to focus on the object. While the interpretation may still be wrong, caused by the image quality or model error, having descriptions of the detected object helped human users to inspect the communication flow and ascertain the point of failure.

Experiments on Yang mod suggested that querying the same agent multiple times for self-reflection and keeping records across images returned sub-optimal performance and significant wastage. With the same agent, only 12 out of 1495 images (1 pole, 11 trees) had different answers between 5 rounds of self-query. The justifications they returned for the same image were very similar, if not identical. Across the three agents, agents only disagreed on 1 image out of all. Asking the same question to a VLM with the same image and prompt was not useful in identifying discrepancies in observations. The lack of variation in answers demonstrated a significant wastage of time and cost in repeating the query to the same very well-conditioned VLM.

Setup	Precision	Recall	Accuracy	F1	False Det	Filter rate
Baseline	0.907 ± 0.000	0.913 ± 0.004	0.952 ± 0.001	0.910 ± 0.002	72	#N/A
Proposed Solution	**0.953 ± 0.001**	0.914 ± 0.000	**0.965 ± 0.000**	**0.933 ± 0.001**	70	30%
Anchoring and feedback (Kwartler mod)	0.936 ± 0.003	0.908 ± 0.004	0.959 ± 0.002	0.922 ± 0.004	66	26%
Debates (Du mod)	0.766 ± 0.000	**0.962 ± 0.000**	0.912 ± 0.000	0.853 ± 0.000	65	29%
Debates with guidance (Lin mod)	0.782 ± 0.015	0.956 ± 0.002	0.917 ± 0.006	0.860 ± 0.009	65	26%
Yang mod	0.813 ± 0.005	0.950 ± 0.005	0.929 ± 0.002	0.876 ± 0.003	106	0%

Table 3 The proposed solution surpasses the baseline and other comparative setups with better F1 score, mainly contributed by eliminating false positives and increasing precision.

5.2 Ablation Study

From asking a plain question to the proposed solution with a detector, reviewer and an aggregator, multiple setups as shown in Section 3.1 were tested and examined. The baseline plain question achieved a high detection score, potentially because the dataset comprised many images of toppled poles or uprooted trees during natural hazards that already appeared on the internet and were pretrained in the models. The differences from the baseline to various ablation setups were analysed.

The setup with a general warmup input prompt returned more “No” answers than the baseline. This increased precision with fewer false positives but decreased recall with more false negatives, resulting in an aggregate drop of F1 by 2.3 percentage points. The additional context may be irrelevant to detecting the toppled poles and caused half-certain first detections to falsely overturn.

In comparison, the self-query setup that provided the first detection as a warmup slightly assisted the final detection, although the improvements were marginal. The 0.5 percentage point increase in F1 score was mainly contributed by overturning 2 missed toppled poles and several false positive tree predictions into true negatives. As opposed to the self-query using the same prompt in Yang mod, the self-query here involved the scope with more qualifications on the targeted attributes. From the communication log, the justifying round provided an alternative interpretation to the image scenes, such as an image showing half a pole collided with by a car was likely toppled overall. The presence of arguments and different query inputs allowed the agent to compare the descriptions of the witnessed scene and infer a potentially different conclusion given the evidence.

The performance of the adversarial setup highly depended on whether the detectors were allowed to take a position freely. The setup with free position was comparable to the baseline setup but worse than the proposed solution, potentially contributed by treating two detectors with equal standing without a hierarchy of reasoning and correcting misperception. This was evidenced by the inferior precision of the proposed solution. This concurred with the observation in Du mod where debates encouraged a corroborated 'Yes' response, leading to some false positives. The setup that mandated an opposite argument however yielded significantly worse F1 (0.704 vs 0.910), mainly contributed by a drastic drop of precision (0.551 vs 0.907). As the decision relied on texts without assigning a plausibility to the argument or the weight of attack towards either argument, the aggregator may take either of the conclusions with an equal weighting. While adversarial setups may theoretically be possible, it had to operate with reliable quantified metrics on the argument quality and quantified metrics on attack or undercuts. Attaining a representative metric for the work will be left for future work.

Setup	Precision	Recall	Accuracy	F1	False Det	Filter rate
Baseline	0.907 ± 0.000	0.913 ± 0.004	0.952 ± 0.001	0.910 ± 0.002	72	#N/A
General warmup	**0.979 ± 0.000**	0.811 ± 0.000	0.945 ± 0.000	0.887 ± 0.000	82	#N/A
Self-query	0.912 ± 0.000	0.917 ± 0.000	0.955 ± 0.000	0.915 ± 0.000	70	6%
Adversarial (free position)	0.893 ± 0.000	0.929 ± 0.000	0.952 ± 0.000	0.911 ± 0.000	70	10%
Adversarial (fixed opposite)	0.551 ± 0.000	**0.972 ± 0.000**	0.782 ± 0.000	0.704 ± 0.000	70	#N/A*
Proposed Solution	0.953 ± 0.001	0.914 ± 0.000	**0.965 ± 0.000**	**0.933 ± 0.001**	70	30%

Table 4 The proposed solution outperformed other ablation setups, translating the use of justification for reviewing answers into strengths in F1 score and the filter rate.
* Despite 36 filter successes, 325 responses adopted the wrong answers from artificial arguments

Setup	Time (s/image)	Cost (£/trial)
Baseline	1.1	0.27
General warmup	2.4	0.83
Self-query	3.1	0.40
Adversarial (free position)	3.3	0.91
Adversarial (fixed opposite)	3.7	0.80
Proposed Solution	2.7	0.62

Table 5 The time and cost incurred on each of the ablation setups

5.3 Comparative study with the general dataset

Experiments on the PhD-base subset showed the proposed solution improved the performance on aggregate, predominantly with the 'no questions'. The proposed solution improved the F1 score by 3.3 percentage points and the recall by 6.4 percentage points, as shown in Table 7. Given the popularity of the COCO dataset that the PhD-base dataset was based on, VLMs may have seen and memorised the objects, leading to a high baseline detection performance. Despite a high benchmarking baseline and potential annotation errors, the proposed solution achieved comparable or greater reduction in false detections than other competing setups. The overturned detections on 'no questions' were mainly false-positive initial detections, whereas the compromise of recall on 'yes questions' was mainly on having false negatives in the initial detections.

Other comparative setups provided similar performance to the proposed setup but cost more time and money. The debating setup without guidance (Du mod) provided similar detection performance but with 75% more time and 38% more cost on 'yes questions', and almost 3 times the time and cost on 'no questions' than the proposed solution as suggested in Table 8. Mixed performances were observed between the debating setups, where the setup with guidance (Lin mod) outperformed the setup without (Du mod) in 'yes questions' and the reverse was shown in 'no questions'. Providing an object for the dissenting agent as guidance to find the object enabled some false negatives to be recovered, improving the recall on 'yes questions'. The debating setup without constraint also incurred double the time and money on 'no questions' than 'yes questions'. The communication log revealed that 241 images (12%) in 'no questions' failed to reach a consensus after 10 rounds of debates, during which the detector or the debater provided no insight.

The self-consistency and voting setup (Yang mod) consumed significantly more time and cost and did not return superior performance. The setup spent about 79 seconds per image and £15 per trial run over 2000 images on both 'yes' and 'no' questions, likely indicating that the setup was not suitable for practical applications. Despite incurring high cost and time, the setup performed worse than the baseline on 'yes questions' and only yielded approximately 1 percentage point gain in F1 from the proposed solution on 'no questions'. Concurring with the observations in the domain-specific dataset, the multiple runs reported a small number (yes: 64-89, no: 34-53, both out of 2000) of questions with varying answers across the 5 rounds of self-query, and even fewer (yes: 6-13, no: 5-12) disagreements across agents. The few variations suggested that the current generation of VLMs would have provided precise, although occasionally wrong answers, but neither self-consistency nor voting were good cures to rectify these. Experiments on the general dataset supported the observations on the domain-specific dataset that self-consistency and voting did not return great improvements that justified the significantly higher cost and time.

5.4 Discussions

The results sections showed the proposed solution completed experiments in both domain-specific and general datasets with substantial time and cost saving than other comparative setups. The time and cost savings may likely be harnessed by negating

Setup	Precision	Recall	Accuracy	F1	False Det	Filter rate
Baseline	0.980 ± 0.000	0.892 ± 0.000	0.881 ± 0.000	0.934 ± 0.000	239	#N/A
Proposed Solution	**0.984 ± 0.000**	0.864 ± 0.000	0.858 ± 0.000	0.920 ± 0.000	280	16%
Anchoring and feedback (Kwartler mod)	0.981 ± 0.000	0.879 ± 0.000	0.870 ± 0.000	0.927 ± 0.000	334	58%
Debates (Du mod)	**0.984 ± 0.001**	0.870 ± 0.002	0.864 ± 0.002	0.924 ± 0.001	280	30%
Debates with guidance (Lin mod)	0.980 ± 0.001	**0.894 ± 0.006**	**0.882 ± 0.006**	**0.935 ± 0.003**	280	30%
Yang mod	**0.984 ± 0.002**	0.857 ± 0.011	0.851 ± 0.008	0.916 ± 0.005	284	1%

Table 6 The proposed solution provided comparable detection performance on the **'yes questions'** of the PhD-base attribute subset.

Setup	Precision	Recall	Accuracy	F1	False Det	Filter rate
Baseline	**0.993 ± 0.000**	0.834 ± 0.000	0.836 ± 0.000	0.907 ± 0.000	328	#N/A
Proposed Solution	0.986 ± 0.000	0.898 ± 0.000	0.890 ± 0.000	0.940 ± 0.000	373	44%
Anchoring and feedback (Kwartler mod)	0.988 ± 0.000	0.877 ± 0.000	0.873 ± 0.000	0.929 ± 0.000	374	37%
Debates (Du mod)	0.990 ± 0.001	0.901 ± 0.005	0.897 ± 0.005	0.943 ± 0.003	373	51%
Debates with guidance (Lin mod)	0.991 ± 0.001	0.867 ± 0.003	0.866 ± 0.004	0.925 ± 0.002	383	38%
Yang mod	0.981 ± 0.002	**0.922 ± 0.006**	**0.908 ± 0.005**	**0.950 ± 0.003**	151	2%

Table 7 The proposed solution substantially improved recall on the baseline and provided roughly the joint highest performance on the **'no questions'** of the PhD-base attribute subset.

	Domain-specific Dataset		General Dataset			
			Yes Questions		No Questions	
Setup	**Time (s/image)**	**Cost (£/trial)**	**Time (s/image)**	**Cost (£/trial)**	**Time (s/image)**	**Cost (£/trial)**
Baseline	1.1	0.27	1.4	0.31	1.3	0.31
Proposed Solution	2.7	0.62	4.4	1.41	5.5	1.66
Anchoring and feedback (Kwartler mod)	5.3	0.68	9.4	1.31	8.1	1.55
Debates (Du mod)	7.2	1.17	7.7	1.94	15.5	4.14
Debates with guidance (Lin mod)	7.2	1.12	8.0	1.26	11.6	1.66
Yang mod	51.3	7.54	78.4	14.65	78.9	15.08

Table 8 The proposed solution yielded superior performance with substantially less time and cost on both the domain-specific and the general dataset.

repeated rounds of communications between agents, which were found to corroborate a dominant answer and were unhelpful in correcting falsified answers. Answers were found to be mostly consistent when the temperature was set to zero in modern models, so asking the same model multiple times with multiple chatbots did not bring much benefit. Comparing with the most similar solution with anchoring and feedback, the proposed solution did not need to perform detection 3 times and instead asked a variety of questions (direct detection, review and substantiate detection, and summarise without detection). The different questions lead to answers with more nuance and comprehensiveness, utilising the query counts better.

The efficiency is coupled with performance gains over the baseline and many comparative setups, predominantly in overturning false positives to increase precision in the Urban Issue dataset, or recall in the 'no questions' of the subset of the PhD dataset. Extending from the variety of questions that the proposed solution asked, answering

each question with an explanation exposed more of the rationale, or weakness if hallucinated. The significance of explanations in providing additional contexts is evidenced by the adversarial setups in the ablation study. Although the setups yielded inferior results, particularly the one mandated to support an artificial contradictory stance, the results showed how agents' explanations were so influential that they altered the reasoning and the conclusions. The proposed solution with a hierarchy of detector-reviewer-aggregator was assisted by how the scope for query was prompted, such that the initial detector was tasked with a slightly narrower scope (to find "electrical poles") and later clarified by the reviewer to include utility and traffic poles but not trees. This avoided many blatant false positives, leaving the reviewer agent to dedicate its effort to differentiate marginal cases.

The query scope further led to discussions on questioning styles and argumentation schemes mentioned in Section 3.1. The authors conducted a pilot trial on enforcing Walton's critical questions on circumstantial evidence or memory to challenge 'Yes' answers, and information-seeking questions to reconstruct narratives on 'No' answers (exact prompt see Section A). This pilot trial outperformed the baseline but achieved less improvement than the proposed solution (F1 score for the setup with Critical Questions: 0.925 vs F1 proposed solution: 0.933). While most of the incorrect answers remained wrong (32 poles, 16 trees), the setup with critical questions tended to focus more on challenging the semantics and descriptions by the detector, essentially analogous to a lawyer challenging a witness as if the stance in a testimony will not change. The proposed solution, on the other hand, enjoyed more freedom in navigating different reasoning paths and referring more to the image. In the handful of answers where the two setups differed, the proposed solution successfully rectified more wrong detections (3 poles, 7 trees), whereas the critical question only struck down 4 more false positive tree detections than the proposed solution. On the positive note, the analysis demonstrated how presenting binary classification as arguments enabled human users to inspect the reasoning flow to the answers and how argumentation schemes helped formulate rebuttals. The pilot study showed that shifting from human-constrained schemes to more autonomy in the VLM may allow the agent to more effectively leverage its own internal heuristics to perform reviews better.

While the generation of arguments by LLMs offers clear advantages in terms of fluency and novelty, it exposes its limitations against traditional formal approaches to argumentation that are grounded in transparency and formal verifiability. Logical and rule-based systems, though often less flexible, enable explicit reasoning steps that can be systematically inspected and validated. The objective of this work is not to reconcile these complementary paradigms, but rather to take a step toward such a reconciliation by beginning from the natural-language end of the spectrum and demonstrating the value of an NLP-centred architectural design. To this end we have proposed a minimal and intuitive, court-inspired multi-agent architecture in which agents are assigned distinct and fixed roles. This architecture demonstrates strong performance not only within our domain-specific setting but also across a broader, more general dataset. A promising direction for future research lies in more closely integrating our approach with rule-based approaches, for instance, by using LLM-agents to generate paired outputs consisting of natural-language arguments and their corresponding

formal representations. In such a framework, the formal component could be mechanically verified for properties such as consistency and adherence to premises, while the natural language output would capture the nuances inherent in natural language, in addition to being more readable by human users.

A further implication of the contemporary shift toward LLM-based agent systems is that the primary decision-making mechanism is effectively a "black box", thereby limiting the applicability of much of the earlier work on roles in MAS. Prior research [15, 81] has for example conceptualised role enactment and deactment as activities that are subject to properties such as the mutual exclusivity of active enacted roles and the requirement that an agent must aim to pursue the objectives associated with an enacted role, potentially alongside its own goals. In some cases, the fact that an agent has been correctly designed to adhere to these constraints can be formally verified. By contrast, the opacity of LLM-based agents precludes direct inspection of their internal decision processes. Consequently, alternative methods are required to assess whether an LLM agent has been successfully enacting a given role, for example through systematic analysis of its observable behavioural trace. Such monitoring could help detect undesirable behaviour, such as excessive sycophancy, that may undermine the integrity of our architecture. Techniques drawn from behaviour composition [82] and runtime verification [7, 83] may prove particularly relevant in this context, enabling the detection of execution-time violations and, where necessary, the prevention of non-compliant actions.

6 Conclusions

Typical VQA using a single VLM agent often results in perceptual, semantic and contextual hallucinations, undermining response validity. This research explores contextual VLM hallucinations by answering as an argument with a conclusion and a justification as the premise. Inspired by the distinctive roles of lawyers and the judge in a courtroom, the proposed solution includes a detector, a reviewer and an aggregator.

The solution was evaluated through experiments using a domain-specific dataset of images with toppled poles to address a key requirement in the roads industry, and the generality of the solution was also demonstrated using a subset of the PhD Dataset prepared from COCO2014. The proposed solution offered a filter rate of 30% against false positive detections and yielded a F1 score of 0.933 in the domain-specific dataset, a 2.3 percentage point improvement from the baseline. The solution allows answers to be inspected at every stage of decision-making to understand how each agent adopts or attacks the argument. It also outperforms comparative solutions in processing time and cost, with 50% time saving from the second-best setup. The reviewer effectively removes out-of-scope false positive detections and retains its reviewing capability without corroborating false positive responses during debates. Comparative studies with the general application datasets further suggested that the proposed setup yielded comparable detection performance with substantially less time and cost. For domain-specific applications such as road maintenance, the proposed solution can be further applied to detect different road defects using VQA and fusion with other detection methods. For general applications, the proposed solution provides

an analytical framework in designing agentic AI architectures through argumentation schemes and enhances interpretability by supplementing justifications to binary Yes/No answers. Further work can also be pursued on enhancing transparency and preventing non-compliant actions with formal representations and rule-based systems.

Acknowledgements. The author (P Lam) is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Future Infrastructure and Built Environment: Resilience in a Changing World (FIBE2) [grant number EP/S02302X/1] and sponsored by the National Highways, Costain and Trimble Solutions. This work is supported by the Digital Roads Prosperity Partnership (DR), supported by the UK EPSRC [grant number EP/V056441/1], Costain, National Highways, the University of Cambridge, Department for Transport and Didimi. The use of Gemini models is supported by the Google Cloud Research Credits program with the award GCP19980904 of USD 1000 worth of cloud credits.

The authors gratefully acknowledge the collaboration of all academic and industrial project partners. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the institutes mentioned above.

Declarations

Funding. The authors received the funding for the research as shown in the Acknowledgements section.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A Prompts

This appendix records the prompts used in the experiments.

Unless otherwise stated, the instruction regarding the JSON format for the output answer is in the following form ('cat' being the detection, review or aggregation process for which the Yes/No answers are needed):

```
instruction = """Output the answers in JSON format.
    Use this JSON schema:

    Output = {cat: "<Yes/No>", 'justification': "str"}
    Return: list[Output]"""
```

A.1 Prompts for the Comparative Study on the Domain-specific Dataset

For all setups, the detector is asked this query prompt:

```
ask_q1 = "Does the image contain any toppled electrical poles? Please
    answer with a justification. " + instruction
```

In the proposed setup, the reviewer is instructed to review the scope with the following prompt:

```
scope_sys = "You evaluate whether a detection result remains valid by
    using the provided scope of toppled poles. If the previous answer
    was a 'Yes', critically examine the elements by attacking the
    premise of the answer, the linkage between the justification and
    the statement, and the statement itself like a cross-examination
    in a courtroom. If the previous answer was a 'No', assist the
    answering like how lawyers ask open questions to help witness to
    recall memories without forcing yourself to answer 'Yes'. You are
    given the image that was detected as a context for what the
    justification was about. Do not carry out a fresh detection again,
     just to evaluate on the peculiarities within a defined scope."
scope_q = scope_sys + "Was the kind of toppled poles indicated
    consistent with the scope of toppled poles that I look for? When
    you answer, please also show me the questions you have asked
    yourself when reaching the conclusion. Also show me your reasoning
     with an explanation."
```

```
scope = "The toppled poles I am looking for are mostly electrical,
    utility or traffic poles that have fallen or are leaning
    significantly from their original upright position. Check the
    toppled objects are not other similar objects such as tree trunks.
     Also do not treat other electric pole defects like having bird
    nests or intervined vines as positive. Note there may be multiple
    defects that appear concurrently on the pole, such as a burning or
     damaged pole falling down. Also near another toppled object,
    there may in fact be a real pole toppling, leaning or
    significantly damaged. Are there such toppled poles of what I
    described in the image? Explain your reasoning here."
ask_scope = f"The detection result is {class_ans}. The justification
    is {just_text}." + scope_q
ask_review = ask_scope + f" The scope is {scope}. Please answer the
    question with a justification." + instruction
ask_review = [ask_review, <the_query_image>]
```

With the answers and justifications from the detector and the reviewer, the aggregator performs its function with the following prompt:

```
agg_q = "You are given a list of detection results and justifications
    from agents. If both the {categories} and the 'inscope' field
    provide the same answer, that will be the final answer. If they
    are different however, provide a verdict based on the 'det_just'
    and 'inscope_just' fields in each input JSON object. Provide a
    justification for your final verdict by considering the 'det_just'
     and 'inscope_just' fields in each input JSON object. Summarise
    these justifications into a single string that supports the final
    verdict. If there were dissenting views, the justification should
    also briefly mention these differing opinions."
ask_agg = f"The detection results from individual agents are {
    store_agent_ans}" + agg_q + instructions
```

In the **"Anchoring and feedback"** setup (Kwartler mod), instead of aggregating responses from the detector and the reviewer, the reviewer's response is fed back to the detector for the final answer.

```
ask_refine = f"The previous detection result was reviewed to check if
    it was within the scope. The scope check result and questions
    asked were {just_dict}. Can you now review your previous answer
    based on the scope check result? The question was {ask_q1}." +
    instruction.
ask_refine = [ask_refine, <the_query_image>]
```

Each comparative setup requires some specific prompts to define its role or communication with other agents. The salient prompts are listed here.

In the **"Debates"** setup (Du mod), the detector is prompted as it has been. The debator is first set up with a prompt about its role:

```
debater_sys = f"You are a debater. There is another main detector who
    will detect on the same image that will be shown to you later. The
     scope of the object to be identified is {scope}. This was not
    given in the question to avoid bias and overfitting the response."
```

In subsequent exchanges, the agent knows of the presence of the other agent and is instructed to update the response with:

```
ask_detect = ask_q1 + f"The <debater/main detector> detected the same
    image with justifications in the previous round as {debate_dict}.
    Using the results from other agents as additional information, can
     you give an updated response to the question." + instruction
```

The debater is reinforced with its role as a debater by being asked to `Remember your role {debater_sys}` in every non-concurred follow-up.

In the **"Debates with guidance"** setup (Lin mod), the debater is first set up as above in `debater_sys`. The detector and the debater are then asked to detect the desired object as the detector in the proposed setup. If they both answer 'No', the system is asked to summarise the justifications for both. If at least one of the agents answers 'Yes', the pipeline proceeds to ask the agent with a positive answer (the detector if both answer 'Yes') about the attributes for what it finds. The query image is provided in each subsequent round of query.

```
further_q = f"For the object you found, what is its shape? Where does
    the object appear? Is there something beside the object? You are
    given the same image used before just to remind the context and
    serve as an evidence. Do not carry out fresh detections for
    objects again. Remind what is the object you found in the 'object'
     field, and provide your descriptions and justifications for these
     descriptions in 'justification'." + instruction
```

The other agent will then be confronted with the positive detection and the attributes.

```
ask_counter = f"The other agent answered 'Yes' by finding an object {
    enquire_dict['object']} with the following description: {
    enquire_dict['enq_just']}. Now please answer the following
    questions one by one. Did you find the same mentioned object? What
     is your opinion on whether it is {enquire_dict['object']} or
    something else? Please answer with a justification." + instruction
```

If both agents now agree with the object identified, the pipeline will consolidate the findings of the enquiry.

```
consolidate_stmt = f"After preliminininary detection {detect_dict} and {
    debate_dict}, the agent that returned a positive detection was
    asked further questions. The question was {further_q} and it
    answered {enquire_dict}. The answer was counter-checked by the
    other agent by asking {ask_counter} and the answer was {
    counter_dict}. Please consolidate the entire answer flow for the
    original question {ask_first} into a single answer and
    justification, bearing in mind at least one detection has returned
     positively in the flow."
ask_consolidate = consolidate_stmt + instruction
```

If the other agent finds something different, its counter finding will be fed back to the detector for a re-interview.

```
ask_again = f"The other agent answered {counter_dict['found_same']}
    with object {counter_dict['counter_object']} for the following
    reasons: {counter_dict['counter_just']}. Now please answer the
    following questions one by one. Did you find the same mentioned
    object? What is your opinion on whether it is {counter_dict['
    counter_object']} or something else? Please answer with a
    justification." + instruction
```

After the re-interview, if the `counter_object` is found, the pipeline consolidates the final answer. If the two agents continue to disagree with the finding, the third judge agent steps in with the communication exchange between the two agents and makes the final decision.

```
ask_consolidate = consolidate_stmt + f"Additionally the first agent
    who answered further questions was re-interviewed with {ask_again}
     and answered {again_dict}. Using the provided image, which result
     is more credible? Please provide a consolidated answer and
    justification. Also find out whether the main detector or the
    debater is the dissent." + instruction
```

In the **"self-consistency and voting"** setup (Yang mod), the three agents acted independently to retrieve the answers. Each agent is prompted with the scope at the beginning of self-consistency queries.

```
ask_det = ask_q1 + f"The scope of the objects or artifacts I referred
    to is: {scope}." + reminder_prompt + instruction
```

The `reminder_prompt` will remind the agent if it has dissented from other agents in the previous round. It was one of the implements in the original solution to ensure consistency across agents (along with diminished voting weight when an agent dissents).

```
reminder_prompt = f"Please remember that you have dissented too often.
     You are now reminded to be more careful in your detection. The
    last time you dissented the aggregator made the decision {
    store_ans[-1]} and your answers were {all_agent_ans[-1][agent]}."
```

Each agent will be asked the same question 5 times. Each of the responses is squeezed into an embedding using "text-embedding-005"[84] model to find outliers. The aggregator is used to perform voting. The weighting of the dissenting agent is diminished until a simple majority vote can be found.

A.2 Prompts for the Ablation Study on the Domain-specific Dataset

Some specific prompts were used in the **ablation studies**.

- The **"general warmup"** setup asked for `What can you see from the image?` before asking the main detection question.
- The **"self-query"** setup asked the detection question in the first query. In the second query, the same chatbot is being asked with the reviewer question `ask_review` as above.

In the **adversarial debating** setup, the detector first performs detection as usual. Another chatbot is then established to perform detection in parallel. In the setup with free positions, the query is:

```
"You are an independent detector. The question for you is {ask_q1}."
```

In the setup that mandates to give an opposite position, the query is:

```
"Another agent has detected the following: {class_ans} with the
    justification {just_text}. You are an independent detector that
    has to reach the opposite decision. Please re-examine the image
    and provide your own detection result along with a justification
    that supports this opposing decision. The original question was {
    ask_q1}"
```

The aggregator then critiques the two results and lays the final response. The prompt is different from the one used for general aggregation, as it is tasked explicitly to attack the arguments made by both agents. The prompt also reminds the aggregator of the original detection question to ensure it responds in a way to reconcile a conclusion for the detection and not for something else.

```
agg_q = f"""
```

```
You are given a list of detection results and justifications from two
    agents. {categories} and {[cat+'_second' for cat in categories]}
    correspond to the detection results from the first and second
    agents respectively. The scope of the {categories} I want to find
    is {scope}. Critically examine answers and justifications from the
     two agents ('det_just' and 'second_just' fields in each input
    JSON object) by attacking the premise of their answers, the
    linkage between their justifications and statements, and the
    statements themselves like a cross-examination in a courtroom. Ask
     questions explicitly to aid your attack and reasoning. \
Rely mainly on the detections and arguments of the two agents instead
    of conducting a fresh detection yourself, even the image is
    provided to you for the context. \
Provide the final verdict of {ask_q1}, the questions you asked to
    arrive at your verdict and a justification for your final verdict.
     Summarise these justifications into a single string that supports
     the final verdict. If there were dissenting views, you should
    also mention these differing opinions.
"""
```

A pilot trial was performed on injecting Walton's argumentation schemes and critical questions in the reviewer of the proposed solution. The critical questions pinpoint towards certain directions of query instead of asking for open and closed questions generally. The scoping prompt for the pilot test is as follows:

```
scope_sys = You are tasked to ask questions based on the response by a
     detector and the scope assigned to you. If the previous answer
    was a 'Yes', critically examine the elements by attacking the
    premise of the answer, the linkage between the justification and
    the statement, and the statement itself like a cross-examination
    in a courtroom. For instance, you can use the critical questions
    in Douglas Waltons' Evidence-based argumentation schemes, such as
    when seeing sign A and inferring a conclusion B, one may ask: \\
CQ1 (Strength): How strong is the correlation between the sign (A) and
     the event (B)? \\
CQ2 (Alternative Explanation): Are there other events that could
    account for this sign? \\
CQ3 (Observation): Is the sign itself verified? \\
CQ4 (Original Perception): How clear was the original perception? \\
CQ5 (Bias): Does the detector have bias that may colour the testimony?

If the previous answer was a 'No', assist the answering like how
    lawyers ask open and information-seeking questions to help witness
     to recall memories without forcing yourself to answer 'Yes'. You
    are given the image that was detected as a context for what the
    justification was about. \\
```

```
Q1 (context cue): What are the context (e.g. weather, colour...) when
    the asserted fact was mentioned? \\
Q2 (temporal sequence): what happened immediately before or after the
    asserted fact happened? \\
Q3 (the scope question): Did you not see it, or did you check and see
    it not there?\\
Q4 (the search depth question): How carefully did you look for the
    object? \\
Do not carry out a fresh detection again, just to ask appropriate
    questions surrounding the desired scope for the detector to
    respond."
```

A.3 Prompts for the Comparative Study on the General Dataset

The PhD dataset has several available types of questions. This series of experiments was performed on the subset of object attributes. For each annotated image, there is a 'yes_question' and a 'no_question'. They are generally straightforward questions, such as "Are the trees green in the image?" (yes_question) and "Are the trees brown in the image?" (no_question). To adapt to the use of evidential reasoning and argumentation for image classification, each of the questions is extended by a request for extra attributes so that agents can review and judge the responses. The question for the detector becomes:

```
eic_q = "Focusing on the object detected, can you describe its
    physical attributes, such as its colour, size, orientation and
    texture? What distinguishing features, if any, does it/do they
    have? Include in the answers in the justification."
det_q = item['yes_question'] or item['no_question']
ask_q1 = det_q + eic_q
```

As explained in Section 3.1, unlike the domain-specific dataset that looks for targeted objects with specific attributes, the general dataset can query for anything in an image. In the absence of specific "scopes" to provide, the reviewer in the **proposed solution** and the **anchoring with feedback** setup will first be asked generally to `"Describe the objects in the image."`. The reviewer then reviews the compatibility of objects with the first detector, as sometimes the detector is queried on peripheral objects that do not contradict the reviewer's self-assessment, but sometimes it makes direct contradictions. The reviewer is then tasked to review whether the original detection should be affirmed or whether the pipeline should reasonably arrive at an alternative conclusion. The reasoning that the reviewer goes through ('scope_sys') is similar to that in the domain-specific dataset.

```
past_observation = f"The previous detection result was {class_ans}.
    The justification was {just_text}."
statement2 = <answer from "describe the objects in the image.">
```

```
scope_sys = "I want you to evaluate whether a detection result remains
     valid. If the previous answer was a 'Yes', critically examine the
     elements by attacking the premise of the answer, the linkage
    between the justification and the statement, and the statement
    itself like a cross-examination in a courtroom. If the previous
    answer was a 'No', assist the answering like how lawyers ask open
    questions to help witness to recall memories without forcing
    yourself to answer 'Yes'. You are given the image that was
    detected as a context for what the justification was about. Do not
     carry out detection again, just to evaluate whether the original
    justification conflicts or supports the re-evaluation."
scope_q = scope_sys + "Would you stick to the original conclusion (
    answer Yes), or is the detection ambiguous and would reasonably
    arrive at a different conclusion (answer No)? If so, please
    explain why the original conclusion still stands. If not, please
    justify your finding."
ask_review = f"I want to review your analysis of this image. You said
    just now {statement2}. \
           In another instance, you said {past_observation}. The
               original question asked was {det_q}" + scope_q +
               instruction
```

When the proposed solution aggregates responses from the detector and the reviewer, as the task of the reviewer is adjusted in the general dataset, the prompt for the aggregator changes. It particularly provides that a 'No' from the reviewer may mean the target object is absent, or it disagrees with the detector (which may have flagged a 'No').

```
agg_q = """You are a mediator evaluating a VQA task. Analyze the
    Detector's verdict ('target', 'det_just') against the Reviewer's
    assessment ('inscope', 'inscope_just'), where 'inscope' signifies
    if the Reviewer agrees (Yes) or disagrees (No) with the Detector.
    Your goal is to determine the final truth by weighing the logical
    strength of 'det_just' versus 'inscope_just'; do not reflexively
    default to the Detector's 'No' if the Reviewer provides evidence
    of a false negative. If the Reviewer provides specific evidence
    that the Detector missed the target or misidentified it, favour
    the Reviewer's logic. Even if both agents output "No", determine
    if they are agreeing that the target is absent, or if the Reviewer
     is claiming the Detector's "No" was an error. Determine a final '
    verdict' (Yes/No), identify 'dissent' by listing the 0-based index
     of any agent whose conclusion you reject (or an empty list if
    they concur), and provide a 'justification' string that
    synthesises the evidence and explains the reasoning behind
    favouring one argument over another in the event of a conflict."""
```

```
ask_agg = f"The detection results from individual agents are {
    store_agent_ans}. The original question was {det_q}." + agg_q +
    instruction
```

In the anchoring with feedback setup, instead of querying a third agent, the detector would be asked to refine its original answer.

```
ask_refine = f"The previous detection result was reviewed to check if
    it was within the scope. The scope check result was {just_dict}.
    Can you now refine your previous answer based on the scope check
    result? The question was {ask_q1}." + instruction
```

In other comparative setups, the scope in the domain-specific dataset is replaced by a general statement to look for compatibility. The prompt controlling the reviewer's reasoning ('scope_q') is identical to the one in the proposed solution. Other agent communications within the setups are governed by prompts modelled after the experiments on the domain-specific dataset.

```
"When you receive that statement, you shall check whether the
    statements are compatible. If your statement does not cover all
    the facts, is the additional information compatible?"
```

Appendix B Pilot study on the choice of models

The state-of-the-art Gemini 3 multimodal framework remained in limited preview at the time of testing. As the research involves substantially comparing time and cost against performance gains, experiments were subsequently performed on the best model suite available for production deployment, the Gemini 2.5 suite. The metrics in Table B1 and throughout the paper generally showed that the Gemini 2.5 suite was very capable of completing the image classification tasks. This pilot study compares the three versions of Gemini 2.5, namely Pro, Flash and Flash-lite in a descending order of sophistication.

For the focus on efficiency, the flash-lite model demonstrated superior performance in time and cost in completing detections on the 1495 images in the Urban Issue subset. A surprising finding is that it provided a better balance of precision and recall and led to better F1 scores on aggregate than the other two models. It was likely that the more sophisticated models, which default to performing chain-of-thought reasoning, spent more time thinking and identifying flaws in the detection. Overthinking may either reject more images than it should (as the Flash model with better precision but worse recall) or accept more images than it should (as the Pro model with inferior precision). With superior performance and efficiency, all experiments in this research were subsequently performed on the Flash-lite model.

Setup	**Precision**	**Recall**	**Accuracy**	**F1**	**Time** (s/image)	**Cost** (£/trial)
Gemini-2.5-flash-lite	0.907 ± 0.000	0.913 ± 0.004	0.952 ± 0.001	0.910 ± 0.002	1.1	0.27
Gemini-2.5-flash	0.954 ± 0.000	0.844 ± 0.000	0.948 ± 0.000	0.896 ± 0.000	2.2	0.97
Gemini-2.5-pro	0.875 ± 0.012	0.916 ± 0.006	0.943 ± 0.004	0.895 ± 0.008	7.2	13.44

Table B1 The flash-lite model provided better F1-score and was more efficient in time and cost on the domain-specific dataset than the more sophisticated models.

References

[1] Fan, H., Liu, C., Janvisloo, N.E., Bian, S., Fuh, J.Y.H., Lu, W.F., Li, B.: MaV-iLa: Unlocking new potentials in smart manufacturing through vision language models. Journal of Manufacturing Systems **80**, 258–271 (2025) https://doi.org/10.1016/j.jmsy.2025.02.017

[2] Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: a review. Frontiers Media SA (2024). https://doi.org/10.3389/frai.2024.1430984

[3] Marie d'Avigneau, A., Potseluyko, L., Anvo, N.R., Taha, H.M., Reja, V.K., Davletshina, D., Lam, P., Silva, L., Al-Tabbaa, A., Brilakis, I.: CAMHighways: The Cambridge Highways dataset. Advanced Engineering Informatics **64**, 103036 (2025) https://doi.org/10.1016/j.aei.2024.103036

[4] Google: What are AI hallucinations? (2025). https://cloud.google.com/discover/what-are-ai-hallucinations?hl=en

[5] Google: What is Retrieval-Augmented Generation (RAG)? (2025). https://cloud.google.com/use-cases/retrieval-augmented-generation?hl=en

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 24824–24837. Curran Associates Inc., New Orleans, LA, USA (2022). http://arxiv.org/abs/2201.11903

[7] Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L., Murthy, A.: Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org, Vienna, Austria (2024). https://doi.org/10.5555/3692070.3692991

[8] Bex, F.J., Van Koppen, P.J., Prakken, H., Verheij, B.: A hybrid formal theory of arguments, stories and criminal evidence. Artificial Intelligence and Law **18**(2), 123–152 (2010) https://doi.org/10.1007/s10506-010-9092-x

[9] Baroni, P., Toni, F., Verheij, B.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games: 25 years later. IOS Press (2020). https://doi.org/10.3233/AAC-200901

[10] Atkinson, K., Bench-Capon, T.: Argumentation schemes in AI and Law. Argument and Computation **12**(3), 417–434 (2021) https://doi.org/10.3233/AAC-200543

[11] Walton, D.: Introducing Some Basic Concepts and Tools. In: Methods of Argumentation, pp. 1–26. Cambridge University Press (2014). https://doi.org/10.1017/cbo9781139600187.001

[12] Modgil, S., Prakken, H.: The ASPIC + framework for structured argumentation: A tutorial. Argument and Computation **5**(1), 31–62 (2014) https://doi.org/10.1080/19462166.2013.869766

[13] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024) https://doi.org/10.1162/tacl.a.00638

[14] Panisson, A.R., Bordini, R.H.: Argumentation Schemes in Multi-agent Systems: A Social Perspective. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10738 LNAI, pp. 92–108. Springer (2018). https://doi.org/10.1007/

978-3-319-91899-0_6

[15] Dastani, M., Dignum, V., Dignum, F.: Role-Assignment in Open Agent Societies. In: Proceedings of the International Conference on Autonomous Agents, vol. 2, pp. 489–496. Association for Computing Machinery (ACM) (2003). https://doi.org/10.1145/860575.860654

[16] Wei, T., Li, T.-W., Liu, Z., Ning, X., Yang, Z., Zou, J., Zeng, Z., Qiu, R., Lin, X., Fu, D., Li, Z., Ai, M., Zhou, D., Bao, W., Li, Y., Li, G., Qian, C., Wang, Y., Tang, X., Xiao, Y., Fang, L., Liu, H., Tang, X., Zhang, Y., Wang, C., You, J., Ji, H., Tong, H., He, J.: Agentic Reasoning for Large Language Models (2026). http://arxiv.org/abs/2601.12538

[17] Cambridgeshire County Council: Street lighting (2025). https://www.cambridgeshire.gov.uk/residents/travel-roads-and-parking/roads-and-pathways/streetlighting

[18] Liu, J., Fu, Y., Xie, R., Xie, R., Sun, X., Lian, F., Kang, Z., Li, X.: PhD: A ChatGPT-Prompted Visual hallucination Evaluation Dataset. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Nashville, TN, USA (2025). https://github.com/jiazhen-code/PhD

[19] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV 2014. Lecture Notes in Computer Science, Vol 8693, pp. 740–755. Springer (2014)

[20] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., Schmidhuber, J.: MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In: ICLR 2024. arXiv preprint (2023). http://arxiv.org/abs/2308.00352

[21] Eghbali, A., Pradel, M.: De-Hallucinator: Mitigating LLM Hallucinations in Code Generation Tasks via Iterative Grounding (2024). http://arxiv.org/abs/2401.01701

[22] Chu, K., Chen, Y.-P., Nakayama, H.: Cohesive Conversations: Enhancing Authenticity in Multi-Agent Simulated Dialogues. Computing Research Repository (2024)

[23] Li, H., Quan Chong, Y., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., Sycara, K.: Theory of Mind for Multi-Agent Collaboration via Large Language Models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 180–192. Association for Computational Linguistics (2023)

[24] Chen, H., Ji, W., Xu, L., Zhao, S.: Multi-Agent Consensus Seeking via Large Language Models (2023). http://arxiv.org/abs/2310.20151

[25] Yu, C.-E., Jalaian, B., Bastian, N.D.: Mitigating Large Vision-Language Model Hallucination at Post-hoc via Multi-agent System. Technical report, AAAI Fall Symposium Series (FSS-24) (2024). www.aaai.org

[26] Jiang, N., Kachinthaya, A., Petryk, S., Gandelsman, Y.: Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations. In: ICLR 2025. arXiv preprint (2024). http://arxiv.org/abs/2410.02762

[27] Liang, X., Yu, J., Mu, L., Zhuang, J., Hu, J., Yang, Y., Ye, J., Lu, L., Chen, J., Hu, H.: Mitigating Hallucination in Visual-Language Models via Re-Balancing Contrastive Decoding. In: Pattern Recognition and Computer Vision, PRCV 2024, pp. 482–496. Springer, Urumqi, China (2025). https://doi.org/10.1007/978-981-97-8620-6_33 . http://arxiv.org/abs/2409.06485

[28] Manevich, A., Tsarfaty, R.: Mitigating Hallucinations in Large Vision-Language Models (LVLMs) via Language-Contrastive Decoding (LCD). In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 6008–6022. Association for Computational Linguistics (2024). http://arxiv.org/abs/2408.04664

[29] Mao, S., Zhang, C., Cai, W.: Through the Magnifying Glass: Adaptive Perception Magnification for Hallucination-Free VLM Decoding (2025). http://arxiv.org/abs/2503.10183

[30] Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-Modal Hallucination Control by Visual Information Grounding. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14303–14312. IEEE Computer Society, Seattle, WA, USA, 2024 (2024)

[31] Wang, C., Fu, W., Zhou, Y.: TPC: Cross-Temporal Prediction Connection for Vision-Language Model Hallucination Reduction (2025). http://arxiv.org/abs/2503.04457

[32] Gao, H., Qu, J., Tang, J., Bi, B., Liu, Y., Chen, H., Liang, L., Su, L., Huang, Q.: Exploring Hallucination of Large Multimodal Models in Video Understanding: Benchmark, Analysis and Mitigation (2025). http://arxiv.org/abs/2503.19622

[33] Ghosh, S., Evuru, C.K.R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., Manocha, D.: Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs. In: ICLR 2025. arXiv preprint (2024). http://arxiv.org/abs/2405.15683

[34] Zou, X., Wang, Y., Yan, Y., Huang, S., Zheng, K., Chen, J., Tang, C., Hu, X.: Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models (2024). http://arxiv.org/abs/2410.03577

[35] Lin, Z., Niu, Z., Wang, Z., Xu, Y.: Interpreting and Mitigating Hallucination in

MLLMs through Multi-agent Debate (2024). http://arxiv.org/abs/2407.20505

[36] Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization (2023). http://arxiv.org/abs/2311.16839

[37] Lyu, X., Chen, B., Gao, L., Song, J., Shen, H.T.: Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp. 122811–122832. Curran Associates, Inc. (2024). https://github.com/BT-C/HIO

[38] Duan, J., Kong, F., Cheng, H., Diffenderfer, J., Kailkhura, B., Sun, L., Zhu, X., Shi, X., Xu, K.: TruthPrInt: Mitigating LVLM Object Hallucination Via Latent Truthful-Guided Pre-Intervention (2025). http://arxiv.org/abs/2503.10602

[39] Villa, A., Alcázar, J.L., Alfarra, M., Araujo, V., Soto, A., Ghanem, B.: EAGLE: Enhanced Visual Grounding Minimizes Hallucinations in Instructional Multimodal Models (2025). http://arxiv.org/abs/2501.02699

[40] Wu, J., Liu, Q., Wang, D., Zhang, J., Wu, S., Wang, L., Tan, T.: Logical Closed Loop: Uncovering Object Hallucinations in Large Vision-Language Models. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 6944–6962. Association for Computational Linguistics (2024). https://github.com/Hyperwjf/LogicCheckGPT.

[41] Klein, L., Amara, K., Lüth, C., Strobelt, H., El-Assady, M., Jäger, P.: Interactive Semantic Interventions for VLMs: A Human-in-the-Loop Investigation of VLM Failure (2024). https://drive.google.com/file/d/1LGZBGOG1wK2-qqdSzqD-esvsEk8qVn_v/preview.

[42] Liu, Y., Ji, T., Sun, C., Wu, Y., Zhou, A.: Investigating and Mitigating Object Hallucinations in Pretrained Vision-Language (CLIP) Models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 18288–18301. Association for Computational Linguistics (2024). http://arxiv.org/abs/2410.03176

[43] Gunjal, A., Yin, J., Bas, E., Ai, S.: Detecting and Preventing Hallucinations in Large Vision Language Models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence, pp. 18135–18143. Association for the Advancement of Artificial Intelligence (2024). www.aaai.org

[44] Shi, J., Zhao, J., Wu, X., Xu, R., Jiang, Y.H., He, L.: Mitigating reasoning hallucination through Multi-agent Collaborative Filtering. Expert Systems with Applications **263** (2025) https://doi.org/10.1016/j.eswa.2024.125723

[45] Sun, X., Li, J., Zhong, Y., Zhao, D., Yan, R.: Towards Detecting LLMs Hallucination via Markov Chain-based Multi-agent Debate Framework. In: ICASSP

2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2025). https://doi.org/10.1109/ICASSP49660.2025.10889448 . https://ieeexplore.ieee.org/document/10889448/

[46] Yang, Y., Ma, Y., Feng, H., Cheng, Y., Han, Z.: Minimizing Hallucinations and Communication Costs: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents. Applied Sciences (Switzerland) **15**(7) (2025) https://doi.org/10.3390/app15073676

[47] Manakul, P., Liusie, A., Gales, M.J.F.: 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, pp. 9004–9017. Association for Computational Linguistics (2023). http://arxiv.org/abs/2303.08896

[48] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: ICLR 2023. arXiv preprint (2022). http://arxiv.org/abs/2203.11171

[49] Liu, L., Yang, D., Zhong, S., Suma, K., Tholeti, S., Ding, L., Zhang, Y., Gilpin, L.H.: Right this way: Can VLMs Guide Us to See More to Answer Questions? In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp. 132946–132976. Curran Associates, Inc. (2024). https://github.com/LeoLee7/Directional_guidance.

[50] Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving Factuality and Reasoning in Language Models through Multiagent Debate. In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org (2024). http://arxiv.org/abs/2305.14325

[51] Kim, B., Shin, W., Lee, K., Jung, Y., Seo, S.: Make VLM Recognize Visual Hallucination on Cartoon Character Image with Pose Information. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5398–5407. IEEE Computer Society (2025). http://arxiv.org/abs/2403.15048

[52] Prabhu, V., Purushwalkam, S., Yan, A., Zhang, J., Xiong, C., Xu, R.: Trust but Verify: Reliable VLM evaluation in-the-wild with program synthesis. In: Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models (2024). https://prove-explorer.netlify.app/.

[53] McDonald, D., Papadopoulos, R., Benningfield, L.: Reducing LLM Hallucination Using Knowledge Distillation: A Case Study with Mistral Large and MMLU Benchmark (2024). https://doi.org/10.36227/techrxiv.171665607.76504195/v1 . https://www.techrxiv.org/users/784557/articles/947452-reducing-llm-hallucination-using-knowledge-distillation-a-case-study-with-mistral-large-and-mmlu-b commit=3eeef7dfb9159c62eeeebc0465d95fa1b0747178

[54] Nguyen, H., He, Z., Gandre, S.A., Pasupulety, U., Shivakumar, S.K., Lerman, K.: Smoothing Out Hallucinations: Mitigating LLM Hallucination with Smoothed Knowledge Distillation (2025). http://arxiv.org/abs/2502.11306

[55] Qu, X., Sun, J., Wei, W., Liu, D., Dong, J., Cheng, Y.: Look, Compare, Decide: Alleviating Hallucination in Large Vision-Language Models via Multi-View Multi-Path Reasoning. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 4428–4441. Association for Computational Linguistics (2025). https://github.com/GasolSun36/MVP.

[56] Zhao, L., Deng, Y., Zhang, W., Gu, Q.: Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance. In: Proceedings of the 42nd International Conference on Machine Learning, vol. 267. Vancouver, Canada (2025). https://github.com/

[57] Sriramanan, G., Bharti, S., Sadasivan, V.S., Saha, S., Kattakinda, P., Feizi, S.: LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp. 34188–34216. Curran Associates, Inc. (2024)

[58] Gosmar, D., Dahl, D.A.: Hallucination Mitigation using Agentic AI Natural Language-Based Frameworks (2025). http://arxiv.org/abs/2501.13946

[59] Kwartler, T., Berman, M., Aqrawi, A.: Good Parenting is all you need Multi-agentic LLM Hallucination Mitigation (2024)

[60] Jennings, N.R., Faratin, P., Lomuscio, A.R., Parsons, S., Wooldridge, M., Sierra, C.: Automated Negotiation: Prospects, Methods and Challenges. Group Decision and Negotiation **10**, 199–215 (2001) https://doi.org/10.1023/A:1008746126376

[61] Walton, D., Reed, C., Macagno, F.: Basic Tools in the State of the Art. In: Argumentation Schemes, pp. 7–42. Cambridge University Press (2012). https://doi.org/10.1017/cbo9780511802034.002

[62] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**, 321–357 (1995)

[63] Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. International Journal of Approximate Reasoning **105**, 252–286 (2019) https://doi.org/10.1016/j.ijar.2018.11.019

[64] Baroni, P., Rago, A., Toni, F.: How Many Properties Do We Need for Gradual Argumentation? In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18) (2018). www.aaai.org

[65] Cocarascu, O., Rago, A., Toni, F.: Explanation via Machine Arguing. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12258 LNCS, pp. 53–84. Springer (2020). https://doi.org/10.1007/978-3-030-60067-9_3

[66] Mayer, T., Cabrio, E., Villata, S.: Transformer-based Argument Mining for Healthcare Applications. In: The 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain (2020). https://gitlab.com/tomaye/

[67] Ruiz-Dolz, R., Heras, S., García-Fornes, A.: An introduction to computational argumentation research from a human argumentation perspective. Autonomous Agents and Multi-Agent Systems **39**(1) (2025) https://doi.org/10.1007/s10458-025-09692-x

[68] Chen, Z., Verdi Do Amarante, D., Donaldson, J., Jo, Y., Park, J.: Argument Mining for Review Helpfulness Prediction. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 8914–8922 (2022). www.joonsuk.org.

[69] Ruiz-Dolz, R., Heras, S., García-Fornes, A.: Automatic Debate Evaluation with Argumentation Semantics and Natural Language Argument Graph Networks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 6030–6040 (2023). https://doi.org/10.5281/zenodo.6531487 . https://doi.org/10.5281/

[70] Cocarascu, O., Stylianou, A., Cyras, K., Toni, F.: Data-empowered argumentation for dialectically explainable predictions. In: Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2449–2456. IOS Press BV (2020). https://doi.org/10.3233/FAIA200377

[71] Reed, C., Long, D.: Content Ordering in the Generation of Persuasive Discourse. In: Proceedings of the 15th International Joint Conference on Artifical Intelligence - Volume 2, pp. 1022–1027 (1997)

[72] Zukerman, I., Mcconachy, R., Korb, K.B.: Using Argumentation Strategies in Automated Argument Generation. In: Proceedings of the 1st International Conference on Natural Language Generation - Volume 14, pp. 55–62 (2000). https://doi.org/10.3115/1118253.1118262

[73] Gretz, S., Bilu, Y., Cohen-Karlik, E., Slonim, N.: The workweek is the best time to start a family - A Study of GPT-2 Based Claim Generation. In: Findings of the Association for Computational Linguistics : EMNLP 2020, pp. 528–544 (2020). www.appen.com

[74] Akhtar, M., Subedi, N., Gupta, V., Tahmasebi, S., Cocarascu, O., Simperl, E.:

ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 13921–13937 (2024). https://github.com/

[75] Walton, D., Reed, C., Macagno, F.: A User's Compendium of Schemes. In: Argumentation Schemes, pp. 308–346. Cambridge University Press (2012). https://doi.org/10.1017/cbo9780511802034.010

[76] Sartor, G.: Argumentation in AI and law (2025). Sartor,Giovanni, ArgumentationinAIandLaw(July25,2025).AvailableatSSRN:https://ssrn.com/abstract=5366436orhttp://dx.doi.org/10.2139/ssrn.5366436

[77] Du, Y., Guo, H., Zhou, K., Xin Zhao, W., Wang, J., Wang, C., Cai, M., Song, R., Wen, J.-R.: What Makes for Good Visual Instructions? Synthesizing Complex Visual Reasoning Instructions for Visual Instruction Tuning. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 8197–8214. Association for Computational Linguistics (2025). https://github.com/RUCAIBox/ComVint.

[78] Google Deepmind: Gemini 2.5 Flash-Lite (2025). https://deepmind.google/models/gemini/flash-lite/

[79] Kahneman, D.: Anchors. In: Thinking, Fast and Slow, pp. 121–123. Penguin (2011). Chap. 11

[80] Himan, A., Randula, D., UOM_220262B: Urban Issues Dataset (2025). https://www.kaggle.com/datasets/akinduhiman/urban-issues-dataset

[81] Dastani, M., Riemsdijk, M., Hulstijn, J., Dignum, F., Ch Meyer, J.-J.: Enacting and Deacting Roles in Agent Programming. In: AOSE 2004, Lecture Notes in Computer Science Volume 3382, pp. 189–204. Springer (2005)

[82] De Giacomo, G., Patrizi, F., Sardiña, S.: Automatic behavior composition synthesis. Artificial Intelligence **196**, 106–142 (2013) https://doi.org/10.1016/j.artint.2012.12.001

[83] Bensalem, S., De Silva, L., Ingrand, F., Yan, R.: A Verifiable and Correct-by-Construction Controller for Robot Functional Levels. Journal of Software Engineering for Robotics **2**(1), 1–19 (2011)

[84] Google Cloud: Text embeddings API (2026). https://docs.cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api