Technical Summary: VALOR-EVAL: Holistic Coverage and Faithfulness Evaluation of Large Vision-Language Models

Kai-Yu Lu

1 Research Problem and Motivation

Large Vision Language Models (LVLMs) integrate a vision encoder and a large language model in order to perform image understanding and natural language generation in a unified framework. Typical LVLM applications include image captioning, visual question answering, and multi modal dialogue. Despite impressive performance, LVLMs are known to suffer from *hallucination*, where the model generates plausible but factually incorrect descriptions that are not grounded in the input image.

Existing hallucination evaluation methods for LVLMs exhibit two main limitations. First, most benchmarks focus almost exclusively on *object existence* hallucinations, such as mentioning a dog that is not present in the image, while largely neglecting hallucinations in *attributes* (such as incorrect color or count) and *relations* (such as incorrect spatial or comparative relations). Second, popular automatic metrics, such as CHAIR for object hallucination, rely on a fixed vocabulary and can only operate in a closed set of object names. This makes them unsuitable for modern open vocabulary captioning, where LVLMs can generate free form natural language, including synonyms and more abstract expressions.

Furthermore, prior work typically emphasizes *faithfulness* to the image, namely avoiding hallucinated content, but does not explicitly encourage *coverage*. A model can achieve high faithfulness by producing extremely conservative and underspecified captions that omit many salient details. In practice, users require captions that are both accurate and informative. The tension between these two objectives is not captured by existing benchmarks and metrics.

The paper summarized in this document addresses these gaps by proposing a new benchmark and a new evaluation framework that jointly quantify hallucinations across objects, attributes, and relations, and that explicitly model the trade off between faithfulness and coverage in open vocabulary generation for LVLMs.[1]

2 Related Work

2.1 Hallucination Benchmarks for LVLMs

Several benchmarks have been proposed to measure hallucination in LVLMs. Many of them primarily target object existence hallucinations. Examples include POPE, NOPE, MERLIM, and other object centric suites where images are paired with prompts that ask whether certain objects exist. These benchmarks typically assume a fixed set of object categories and constrain the evaluation to a discriminative decision about presence or absence.

Other works have started to broaden the scope of hallucination evaluation. Some benchmarks attempt to include simple attributes or relations, or to adopt free form prompts rather than fixed question templates.

[1] The acronym VALOR stands for vision language attribute, relation, and object coverage and faithfulness.

Nevertheless, these efforts still tend to emphasize one specific dimension, such as object existence or very limited attribute types, and they rarely perform systematic analysis across objects, attributes, and relations in a unified setting.

2.2 Evaluation Metrics and CHAIR

The CHAIR metric is a widely used automatic measure for object hallucination in image captioning. It operates on captions by extracting object mentions from a caption and comparing them against ground truth object annotations in an image. CHAIR is defined over a fixed vocabulary, for example the eighty object categories in the MS COCO dataset. This design implies that CHAIR cannot directly handle open vocabulary expressions or synonyms that fall outside the predefined list. In addition, CHAIR focuses mainly on hallucinated objects and does not explicitly model how much of the ground truth content a caption covers.

Subsequent works have proposed automatic evaluation models or LLM based scoring protocols that aim to replace human judgements, but these systems often introduce their own biases or remain restricted to closed vocabularies and object level evaluation.

2.3 Positioning of VALOR-EVAL

The VALOR-EVAL framework generalizes CHAIR in two key directions. First, it extends the evaluation scope to three categories of hallucinations: object existence, attributes, and relations, with further subdivision of attribute and relation types. Second, it replaces fixed vocabulary matching with a large language model based semantic matching procedure that operates at the feature level in an open vocabulary setting.

The accompanying benchmark VALOR-BENCH provides human annotated ground truth across the three feature types using carefully selected images that are designed to expose associative biases in LVLMs. Together, VALOR-BENCH and VALOR-EVAL form a comprehensive evaluation suite for generative hallucinations of LVLMs.

3 Dataset Construction

3.1 Feature Definitions

The benchmark VALOR-BENCH evaluates hallucinations on three principal types of features in an image.

Object existence. Object existence features are pairs of visual entities that appear in an image, namely *object object* combinations. Both foreground and background objects are included. For example, an image showing a lady sitting on a bench in front of a building would have objects such as “lady”, “bench”, and “building”.

Attributes. Attribute features describe visual properties of objects and are modeled as *object attribute* pairs. The benchmark focuses on two attribute dimensions, namely color and counting, and splits attributes into two subcategories:

- Object attributes, such as the color and count of non human objects. An example is “six green apples on the table”, corresponding to attributes (green, apple) and the count of apples.
- People attributes, where each person is annotated with clothing colors and the total number of people. For example, “a woman wearing a red jacket and black shoes” is annotated as (woman, (red, jacket), (black, shoes)), together with the total count of people in the image.

Table 1: Overview of VALOR-BENCH subsets, including feature category, sub category, number of images, and source.

Category	Sub category	#Images	Source
Object existence	–	50	GQA
Attribute	Object attributes (color and count)	27	Pexels
Attribute	People attributes (clothes and count)	34	Pexels
Relation	Positional relations	50	GQA
Relation	Comparative size relations	50	GQA

Relations. Relation features are triples describing how objects are related and are modeled as *object relation object* combinations. Two relation types are considered:

- Positional relations, such as left, right, top, and bottom, for example "the bed is to the left of the table" and the inverse "the table is to the right of the bed".
- Comparative relations based on relative size, where objects are ranked from largest to smallest (for example "1. bed, 2. table, 3. cup").

These definitions are consistently used in both the benchmark construction and the evaluation framework.

3.2 Datasets and Sources

The benchmark leverages two main sources of images:

- The GQA dataset, which provides richly annotated real world images with object, attribute, and relation annotations.
- A copyright free stock photo website (Pexels), used to obtain additional images for attribute evaluation where GQA annotations are sparse.

The final VALOR BENCH dataset is structured as summarized in Table 1.

3.3 Co occurrence Statistics

To construct challenging test cases, the benchmark explicitly exploits associative biases in the training data. The core hypothesis is that LVLMs are exposed to frequent co occurrences of certain object, attribute, and relation combinations during training and therefore learn strong statistical associations. When a test image contains only one component of a frequently co occurring combination, the model may hallucinate the missing components.

The first step is to compute co occurrence statistics in GQA. For each object and feature pair, the paper defines a conditional probability:

$$P(\text{feature} \mid \text{object}) = \frac{\mathrm{Frequency}(\text{feature}, \text{object})}{\mathrm{Frequency}(\text{object})}, \tag{1}$$

where the symbol feature ranges over all annotated objects, attributes, and relations in the dataset, the symbol object denotes an annotated object category, and $\mathrm{Frequency}(\text{feature}, \text{object})$ counts how many images contain both the feature and the object. The term $\mathrm{Frequency}(\text{object})$ counts how many images contain the object irrespective of other features. Intuitively, $P(\text{feature} \mid \text{object})$ measures how likely a specific feature is to appear when a given object is present. This conditional probability is used to identify strong and weak associations between objects and features.

Based on these conditional probabilities, five scalar metrics are derived for each object by aggregating over all associated features. These metrics are the maximum conditional probability, the average conditional probability, the difference between maximum and average, the difference between average and minimum, and the difference between maximum and minimum. Together, they highlight objects that have particularly skewed or concentrated feature distributions and therefore strong associative biases.

3.4 Image Selection via Associative Bias

The benchmark construction then uses the co occurrence statistics to select images that are likely to induce hallucinations. Several sets are defined in terms of the conditional probabilities.

First, a set of objects O with pronounced co occurrence dependencies is selected:

$$O = \left\{ \arg\max_{o} P(f \mid o) \mid f \in F \right\}, \tag{2}$$

where F denotes the set of all annotated features in the dataset, the variable o ranges over all annotated objects, and $P(f \mid o)$ denotes any of the previously described statistical dependency measures between feature f and object o. The set O therefore collects objects that have at least one feature with maximal association, capturing strong co occurrence patterns.

Second, for each object in O, the benchmark identifies a set of minimally associated features:

$$I = \left\{ \arg\min_{i} P(i \mid o) \mid i \in F_o,\ o \in O \right\}, \tag{3}$$

where F_o is the set of features that co occur with object o at least once, and $P(i \mid o)$ is the conditional probability that feature i appears when object o is present. The set I therefore contains features that are weakly associated with their corresponding objects. Images containing such pairs can serve as counter examples to strong biases.

Third, a set of highly associated features is defined:

$$H = \left\{ \arg\max_{h} P(h \mid o) \mid h \in F_o,\ o \in O \right\}, \tag{4}$$

where symbols have the same meaning as in Equation (3). The set H represents strongly co occurring features that LVLMs are likely to hallucinate when their associated object is present.

Finally, a collection of images C is constructed:

$$C = \{c : (o, f) \mid o \in O, f \in I, \text{ and } f \notin H\}, \tag{5}$$

where c denotes an image containing object o together with feature f. The constraint $f \notin H$ ensures that the image does not contain any of the strongly associated features, even though the object is present. Intuitively, the set C contains images that intentionally break strong statistical associations, so that models trained on the original co occurrence patterns are tempted to hallucinate missing features.

These steps are applied separately to construct subsets for object existence and relation features based on GQA annotations. For attribute features, the limited attribute coverage in GQA makes it difficult to find sufficient images, so additional images are retrieved from Pexels using the previously computed statistics as guidance.

3.5 Human Annotation Protocol

All images selected for VALOR BENCH are manually reviewed and annotated by expert annotators. The annotation process follows the feature definitions in Section 3.1 and produces, for each image, a set of ground truth features F_G that includes object existence entries, attribute entries, and relation entries.

For object existence, the annotators verify and extend the original object annotations so that all visually salient entities, including background elements, are covered. For attributes, annotators record color and count for objects and the number of people and clothing colors for human subjects. For relations, annotators record positional relations between pairs of objects and comparative size rankings for all objects in the image.

Each image is represented as a tuple (I, F_G, p_G), where I is the image, F_G is the list of ground truth features, and p_G is a prompt used later to query LVLMs for generation.

4 Query Protocol and Task Definitions

The benchmark defines a query protocol that maps each feature subset to a specific generation task and prompt. All tasks are formulated as open vocabulary captioning where the LVLM must produce a natural language description.

4.1 Object Existence Task

For the object existence subset, the prompt p_G is:

> Write a detailed description of the image. Provide information about all objects in front and background.

The LVLM must generate a caption that lists and describes all objects in the scene. The evaluation focuses on whether the mentioned objects match the ground truth object existence annotations and whether the caption omits or hallucinates objects.

4.2 Object Attribute Task

For object attributes such as color and count, the prompt is:

> Write a detailed description of the image. Provide information about the total number and colors of all objects from left to right and up to bottom.

The LVLM is required to enumerate objects, their counts, and their colors in a structured manner. The evaluation examines whether colors and counts are accurate and whether the model correctly captures the attributes of each object.

4.3 People Attribute Task

For people attributes, the prompt is:

> Write a detailed description of the image. Provide information about the total number of people and colors of clothes for each person from left to right.

The model must recognize the number of human subjects and the clothing colors associated with each individual. Hallucinations may arise when the model fabricates extra people or incorrect clothing details.

4.4 Positional Relation Task

For positional relations between objects, the prompt is:

> Describe the positional relationship between all the objects in the image in detail, using left, right, top, and bottom etc, from the view of the observer.

The model needs to generate relational statements that describe where objects are located relative to each other, for example “the plate is in the center of the table” or “the mouse is to the left of the plate”. The evaluation checks whether these relations match the annotated positional relations.

4.5 Comparative Relation Task

For comparative size relations, the prompt is:

> Rank the size of all the objects in the image in detail, from large to small.

The LVLM must rank objects according to their size in the image. The evaluation compares the generated ranking against the annotated ranks to detect hallucinated size relations or mis orderings.

5 Modeling Approach

5.1 Overview of VALOR-EVAL

The VALOR EVAL framework is a two stage evaluation pipeline that uses a large language model as an auxiliary agent. The first stage extracts structured features from LVLM generated captions, and the second stage matches these extracted features against the ground truth features using semantic reasoning. Both stages operate in an open vocabulary setting.

Concretely, given an image I and a benchmark prompt p_G, an LVLM under evaluation produces a free form response R, which is a natural language caption. A separate large language model, instantiated as GPT 4 in the original paper, is used as an evaluation agent. This agent performs two tasks:

- Feature extraction: given R, the agent extracts a set of features $F_R = \{f_R^1, \ldots, f_R^m\}$ that summarize the objects, attributes, and relations mentioned in the caption.
- Feature matching: given F_R and the ground truth features $F_G = \{f_G^1, \ldots, f_G^n\}$, the agent determines which generated features are correct matches to ground truth features and which generated features are broader but still semantically compatible.

The framework relies on carefully designed prompts for feature extraction and feature matching, which instruct the evaluation agent to output normalized structured representations instead of free text.

5.2 Feature Extraction

Feature extraction uses prompts p_E that are tailored to each feature type. For example, for object existence, the agent is asked to list all objects mentioned in the caption. For attributes, the agent is asked to output tuples of objects and their colors or counts. For relations, the agent is asked to output structured triplets representing positional or comparative relations.

The result of feature extraction is a set F_R of generated features for each caption. Each element f_R^i is a structured representation such as an object name, an object attribute pair, or an object relation triple. This representation abstracts away from the surface form of the caption while preserving its semantic content.

5.3 Feature Matching and Semantic Alignment

Feature matching aligns F_R with the ground truth feature set F_G using the evaluation agent with matching prompts p_M. The matching yields two dictionaries:

- A dictionary of directly matched features D_M. Each entry in D_M pairs a generated feature f_R^i with a ground truth feature f_G^j when the evaluation agent judges them to be semantically equivalent at the chosen level of granularity.
- A dictionary of broader conceptual matches D_B, where each generated feature is judged to be a broader category that still semantically includes the ground truth feature. For instance, the generated attribute $(\text{red}, \text{clothes})$ may be considered a broader match for the ground truth attribute $(\text{red}, \text{dress})$.

Unlike traditional metrics that rely on string matching inside a fixed vocabulary, this semantic matching procedure allows VALOR EVAL to recognize synonyms, hypernyms, and other semantically related expressions. For example, the generated attribute $(\text{plaid}, \text{shirt})$ can be matched to a ground truth attribute $(\text{checkered}, \text{shirt})$, because both describe similar fabric patterns.

5.4 Faithfulness Metric

The faithfulness metric quantifies how accurate a caption is relative to the ground truth features. It considers both direct matches and broader conceptual matches. The metric is defined as

$$\text{Faithfulness}(R, F_G) = \frac{|D_M \cup \text{set}(D_B)|}{|F_R|} \in [0, 1], \tag{6}$$

where R is the generated caption, F_G is the ground truth feature set, F_R is the set of generated features, D_M is the set of directly matched features, and $\text{set}(D_B)$ denotes the set of generated features that have broader but still correct semantic matches in the ground truth.

The numerator $|D_M \cup \text{set}(D_B)|$ counts how many generated features are judged correct, either exactly or at a broader conceptual level. The denominator $|F_R|$ counts the total number of features mentioned by the model. Intuitively, Equation (6) measures the proportion of the model's statements that are supported by the image annotations. A high faithfulness score indicates that the caption rarely hallucinates features, while a low score suggests that many generated features do not correspond to true objects, attributes, or relations in the image.

5.5 Coverage Metric

The coverage metric measures how comprehensively the generated caption describes the image. It is defined as

$$\text{Coverage}(R, F_G) = \frac{|\text{set}(D_M)|}{|F_G|} \in [0, 1], \tag{7}$$

where $\text{set}(D_M)$ is the set of ground truth features that have at least one direct match in the generated feature set, and $|F_G|$ is the total number of ground truth features in the image.

The numerator therefore counts how many ground truth objects, attributes, or relations are successfully captured by the generated caption, while the denominator counts how many should have been captured in principle. A high coverage score indicates that the model mentions most of the salient features of the image, whereas a low coverage score reveals that the model omits many ground truth features and produces under informative captions.

Together, the faithfulness and coverage metrics explicitly characterize the trade off between avoiding hallucinations and providing rich, detailed descriptions.

5.6 Relation to CHAIR and CHAIR_LLM

The original CHAIR metric operates only on objects and relies on a fixed object vocabulary with synonym lists. VALOR EVAL can be viewed as a generalization in which:

- the vocabulary is open and features include objects, attributes, and relations;
- matching is performed by an LLM using semantic reasoning rather than string matching;
- coverage is explicitly accounted for alongside hallucination.

In addition to the full VALOR EVAL framework, the paper introduces an LLM augmented variant of CHAIR, referred to as CHAIR_LLM, where the LLM is used only to improve object extraction and matching for CHAIR while preserving the general form of the original metric.

6 Empirical Results

6.1 Experimental Setup

The evaluation uses VALOR BENCH together with VALOR EVAL to assess ten established LVLMs covering both open source and commercial systems. The evaluated models include InstructBLIP, LLaVA 1.5, MiniGPT 4 v2, mPLUG Owl2, BLIVA, CogVLM, InternLM XComposer2, Qwen VL Chat, Emu2, and GPT 4V. These models represent different design choices in terms of visual encoders, alignment modules, and language backbones, for example EVA or CLIP based vision transformers and Vicuna or other large language models as text backbones.

Each model is prompted using the task specific prompts defined in Section 4. GPT 4 is used as the evaluation agent for feature extraction and matching. The main metrics reported are the faithfulness and coverage scores as defined in Equations (6) and (7), computed per subset and averaged across subsets.

6.2 Coverage Faithfulness Trade off across Models

Table 3 in the original paper reports faithfulness and coverage scores for each model across the five feature subsets and their averages. Several systematic patterns are observed.

The model Emu2 achieves the highest average faithfulness score, approximately 74.98%, indicating that most features it mentions are correct. For example, its faithfulness reaches about 94.2% on object existence and about 87.5% on comparative relations. However, Emu2 exhibits very low coverage, with an average coverage of about 8.1%, and particularly low coverage for attribute people and positional relations. This suggests a very conservative generation strategy where the model prefers to mention only a few highly reliable features, which reduces hallucinations but results in under informative captions.

In contrast, GPT 4V achieves the highest average coverage, about 28.0%, meaning that it describes a wider range of objects, attributes, and relations per image. Its coverage on object existence reaches approximately 38.8%, which is the highest among all models, and its coverage on people attributes is also relatively strong. However, its average faithfulness, around 54.62%, is not the highest, indicating that the greater level of detail comes at the cost of more hallucinated features.

Other models, such as LLaVA 1.5 and CogVLM, occupy intermediate positions. They achieve moderately high faithfulness and coverage values, indicating a more balanced trade off between accuracy and informativeness. The paper highlights LLaVA 1.5 in particular as an example of a model that achieves competitive performance through efficient use of instruction tuning data.

Overall, these results support the paper's claim that different LVLMs implicitly choose different operating points on the faithfulness coverage spectrum. Some models prioritize precision and avoid hallucinations at the cost of incomplete descriptions, while others prioritize richer descriptions at the cost of additional hallucinations. VALOR EVAL exposes these differences quantitatively.

6.3 Correlation with Human Judgements

To validate the reliability of VALOR EVAL, the paper compares the automatic faithfulness and coverage scores against human judgements for captions generated by InstructBLIP on a subset of images. The Pearson correlation coefficient is used to quantify the linear relationship between automatic scores and human ratings.

The correlation results are high across most subsets. For object existence, the correlation between automatic and human faithfulness scores is about 0.91, and the correlation between automatic and human coverage scores is about 0.89. For attributes and comparative relations, the correlation values are even higher, often around 0.98 or 0.99. Positional relations, which involve longer and more complex descriptions, yield somewhat lower but still strong correlations, approximately 0.78 for faithfulness and 0.86 for coverage.

These findings indicate that VALOR EVAL's LLM based scoring is closely aligned with human judgements across multiple feature types, providing evidence that the automatic evaluation is both effective and reliable.

6.4 Effect of Co occurrence Based Data Selection

An ablation study examines how the co occurrence based image selection affects the difficulty of the benchmark. The paper compares model performance on two types of evaluation data:

- images selected randomly from GQA and annotated in the same way as VALOR BENCH;
- images selected using the co occurrence statistics and associative bias method described in Section 3.

For object hallucination, three representative models are evaluated: InstructBLIP, LLaVA 1.5, and GPT 4V. For InstructBLIP, faithfulness decreases from 76.5% on randomly selected images to 74.5% on the co occurrence selected images, while coverage remains similar. For LLaVA 1.5, the drop in faithfulness is more substantial, from 84.5% to 72.1%, and coverage also decreases slightly. GPT 4V also shows a moderate reduction in both faithfulness and coverage when evaluated on the co occurrence selected images.

The paper further notes that co occurrence selected images contain, on average, more objects than randomly selected images. This combination of higher object density and broken co occurrence patterns makes the benchmark more challenging and more sensitive to hallucination tendencies, confirming the value of the proposed data selection method.

6.5 Comparison with Original CHAIR Metric

A second ablation study compares the LLM augmented CHAIR_LLM variant against the original CHAIR metric on a small set of twenty MS COCO images. These images are re annotated in a way that is compatible with CHAIR, using the same synonym lists as the original metric.

Two accuracy scores are reported. The first, denoted as Acc(F), measures how accurately the metric identifies hallucinated objects in captions, penalizing false positives. The second, denoted as Acc(C), measures how many of the objects mentioned in the caption are correctly detected by the metric, effectively evaluating the extraction phase.

Across three models, CHAIR_LLM significantly outperforms the original CHAIR. For InstructBLIP, Acc(F) increases from about 11.11% with CHAIR to about 88.89% with CHAIR_LLM, and Acc(C) increases from about 80.66% to 100%. For LLaVA 1.5, Acc(F) improves from about 30.00% to 90.00%, and

Acc(C) improves from about 83.52% to 97.08%. For GPT 4V, which generates longer and more complex captions, the original CHAIR achieves only about 5.88% Acc(F), whereas CHAIR_LLM boosts this to about 82.35%, and Acc(C) rises from about 82.35% to about 98.17%.

These results demonstrate that LLM based semantic extraction and matching provide a considerable advantage over fixed vocabulary string matching, especially for modern LVLMs that employ diverse and expressive language.

6.6 Implementation and Evaluation Details

The paper evaluates pre trained LVLMs without further fine tuning. Therefore, traditional training hyperparameters such as learning rate, batch size, and optimization method are not central to the study. Instead, the key configuration choices concern:

- the prompts used to query LVLMs, which are fixed for each subset and described in Section 4;
- the use of GPT 4 as a single evaluation agent for both feature extraction and matching;
- the choice of evaluation metrics, namely faithfulness and coverage as defined in Equations (6) and (7), along with Pearson correlation for comparisons with human judgements.

The hardware and software environment is not emphasized in the paper, because the main contribution lies in the benchmark design and evaluation methodology rather than in training new models.

7 Summary

7.1 Key Contributions

The paper makes three primary technical contributions.

First, it introduces VALOR BENCH, a human annotated benchmark designed to evaluate LVLM hallucinations across three feature types: objects, attributes, and relations. The benchmark further decomposes attributes into object and people attributes and relations into positional and comparative relations. Images are selected using co occurrence statistics to expose associative biases and induce challenging hallucination scenarios.

Second, it proposes VALOR EVAL, a two stage LLM based evaluation framework that generalizes the CHAIR metric to open vocabulary settings and to multiple feature types. VALOR EVAL uses a large language model to extract structured features from captions and to semantically match them to ground truth features, then computes faithfulness and coverage scores that quantify both hallucination and informativeness.

Third, it provides an empirical analysis of ten LVLMs on VALOR BENCH, showing that different models occupy different positions in the trade off between faithfulness and coverage and that even advanced models such as GPT 4V still exhibit substantial hallucinations. The paper also demonstrates that VALOR EVAL correlates strongly with human judgements and that an LLM augmented variant of CHAIR significantly improves object hallucination detection.

7.2 Limitations

The paper explicitly acknowledges several limitations. The attribute and relation coverage in VALOR BENCH, while richer than in many prior works, is still restricted to a subset of possible attributes and relations, namely color, counting, positional relations, and comparative size. Other important dimensions such as material, texture, actions, and abstract relations are not yet included.

In addition, the evaluation pipeline relies on GPT 4 as a single evaluation agent. Any biases, preferences, or systematic errors of GPT 4 may propagate into the faithfulness and coverage scores. Although the strong correlation with human judgements suggests that the evaluation is meaningful, the dependence on a proprietary model may limit reproducibility in some settings.

Finally, the benchmark uses a single prompt per subset. Some LVLMs may be more sensitive to prompt wording than others, and additional prompt engineering or diverse prompts could further reveal model behavior.

7.3 Future Directions

Future research directions suggested by the paper include expanding VALOR BENCH to cover a wider range of attributes and relations, incorporating additional datasets beyond GQA and Pexels, and exploring alternative or multiple evaluation agents to reduce dependence on a single LLM. Another promising direction is to use VALOR EVAL as a diagnostic tool for developing training strategies, decoding algorithms, or regularization methods that explicitly optimize both faithfulness and coverage, thereby reducing hallucinations without sacrificing informativeness.