UNIVERSITY OF CALIFORNIA,
IRVINE

Natural Language Expressions of Uncertainty in VLM Product Captions: An Evaluation for Blind and Low-Vision Users

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Informatics

by

Dwayne Morgan

Thesis Committee:
Professor Dr. Anne Marie Piper, Chair
Chancellor's Professor Dr. Gillian Hayes
Lecturer Dr. Alberto Krone-Martins

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To Anne Marie and Kapil: I am truly grateful to both of you for your guidance throughout our time together. I appreciate your continued patience and dedication to pushing me to think more deeply have shaped how I approach problems as a researcher. I have left every meeting reflecting on how to refine my thinking and asking *how* every choice matters. I am grateful to have grown as a researcher during this time and aspire to continue that growth. Thank you both.

To my family: Mom and Dad, thank you for all your support — the care packages, the phone calls (often when I forgot to eat), and the reassurance that made this work possible. Ambar, your endless support and patience will forever be appreciated. The late nights of writing and reading would truly not have been possible without you.

ABSTRACT OF THE THESIS

Natural Language Expressions of Uncertainty in VLM Product Captions: An Evaluation for Blind and Low-Vision Users

By

Dwayne Morgan

Master of Science in Informatics

University of California, Irvine, 2026

Professor Dr. Anne Marie Piper, Chair

Vision-Language Models (VLMs) are increasingly used by blind and low-vision (BLV) people to identify products in everyday life, yet producing error-free captions under real-world image conditions remains an open challenge. Under these conditions, models often produce descriptions that include uncertain language, and as errors are difficult to verify non-visually, how models communicate uncertainty in their output matters as much as whether they are accurate. Hedging language, expressions such as "likely," "appears to be," and "possibly", is a natural way uncertainty manifests in VLM output, but whether VLMs hedge in ways that are actually aligned — whether models express uncertainty where they are likely wrong and withhold it where they are likely right — is unknown. In this thesis, I investigate (1) the extent to which four VLMs (GPT-4.1, Gemini-2.5 Flash, LLama-90B 3.2, and Molmo-72B) produce hedging language in product captions, and (2) whether hedging co-occurs with caption inaccuracy and image quality degradation across 1,859 product images drawn from BLV users. I find that hedging rates vary substantially across models (from under 1% to over 17%) and that hedging shows a weak and model-dependent association with caption accuracy. This work calls for evaluation frameworks that go beyond accuracy to measure uncertainty alignment as well as for training pipelines that treat uncertainty communication as a design objective rather than as emergent artifacts of pre-training data.

1 Introduction

Human communication is full of uncertainty. In everyday interactions, people routinely make sense of tentative or incomplete information by drawing on contextual cues, prior experience, and the ability to independently verify claims. Hedging language, when speakers soften their claims with words like "I think," "probably," or "maybe" rather than stating them outright, is one of these features of conveying uncertainty. This language allows listeners to judge how reliable a claim is, and lets speakers signal what they do and do not know (Lakoff, 1973; Prokofieva and Hirschberg, 2014).

Large language models (LLMs) are capable of producing this same vocabulary. Trained on enormous quantities of human text, modern LLMs routinely generate phrases like "it appears," "possibly," and "I'm not sure" in their outputs. While these expressions match the linguistic uncertainty markers humans use, a model hedge is a token sequence selected because similar sequences appeared in similar contexts during training (Xiong et al., 2023; Mielke et al., 2022). Whether such a hedge suggests anything meaningful about what a model truly "knows" is an open question.

Vision-Language Models (VLMs) used as visual assistants by blind and low-vision (BLV) users are one such setting where hedging language manifests (Be My Eyes, 2023; Meta, 2025). VLM-based applications like Be My AI and Seeing AI now help BLV users identify products, read packaging, and interpret their surroundings. A 2024 survey found that 87% of BLV

respondents already used or would consider using AI as a visual assistant (Karamolegkou et al., 2024). For many BLV users, these VLM-based tools offer autonomy and agency that alternative assistive technology cannot reliably provide. Remote sighted assistance services are subject to volunteer availability and technical failures, and even trained visual interpreters encounter challenges in information acquisition and social engagement (Tang et al., 2025).

However, VLMs are not without errors. VLMs produce inaccurate and hallucinated output that blind users find difficult to verify non-visually. What's more, model performance degrades further on the low-quality images that BLV users commonly capture in real-world conditions (Alharbi et al., 2024; Adnin and Das, 2024; Garg et al., 2026). In consideration of these errors, research has suggested that fully error-free models are not a realistic near-term goal (Lin and Jackson, 2023; Tang et al., 2025). When errors are unavoidable, how a model communicates its uncertainty becomes equally as important as whether it is accurate. When a VLM hedges in a caption ("this is *likely* a can of chicken soup"), a BLV user has limited means to independently verify whether the hedge is meaningful and prior research finds that users interpret uncertainty cues as signals about system reliability, adjusting their reliance accordingly (Kim et al., 2024). The cost of an absent or inaccurate hedge in critical scenarios, such as identifying medication or detecting allergens, can mean the difference between safe use and serious harm.

This thesis investigates how, when, and how reliably VLMs hedge in product-identification captions for BLV users. Prior work in HCI has defined uncertainty interchangeably with instability, risk, and ambiguity (Tang et al., 2025). What is missing is an account of whether the signals models already produce are actually aligned with when the model is likely to be wrong.

I ask three questions:

1. In VLM product captioning, how often do models hedge, and with what vocabulary?

2. Does hedging in model-generated captions reliably indicate lower accuracy?

3. Does hedging correspond to the presence and type of image quality issues?

I address these questions by extending an existing dataset of VLM-generated product captions annotated for accuracy and image quality with hedging annotations across four models (GPT-4.1, Gemini-2.5, Llama-90B, and Molmo-72B) (Garg et al., 2026).

Across the 1,859 product images in the dataset, hedging behavior varies substantially by model: Llama hedges in 17.4% of captions, Molmo in 11.9%, GPT in 6.3%, and Gemini in only 0.8%. Hedging is statistically associated with caption error in three of the four models, but the association is weak. Image quality tells us more about whether a caption is wrong than the model's hedging language does. Furthermore, where hedging *does* respond to image quality, it does so unevenly. Model identity was found to be the strongest predictor of whether a caption hedges at all, suggesting a gap between how uncertainty is internally represented and how it is communicated to users and raising concerns about how users interpret and act on these signals in real-world settings.

This thesis contributes the following:

1. An annotation approach for identifying and categorizing hedging language in open-ended VLM captions, grounded in established linguistic theory (Lakoff, 1973; Prokofieva and Hirschberg, 2014).

2. An extension of an existing product-caption dataset (Garg et al., 2026) that adds hedging annotations alongside the existing accuracy and image quality labels, supporting future work on disability-oriented evaluation of VLM captioning.

3. A cross-model characterization of hedging in VLM-generated product captions, show-

ing that hedging frequency varies by model and is weakly correlated with accuracy in contrast with image quality.

2 Background

This thesis sits at the intersection of three bodies of literature: how models quantify and communicate uncertainty, how users interpret those uncertainty signals from models, and how blind and low-vision people use and navigate the limitations of AI visual assistants.

While substantial progress has been made in developing VLMs for general image captioning, the specific challenges of product identification for BLV users, particularly how linguistic expressions of uncertainty relate to model accuracy, remain underexplored.

2.1 Quantifying Uncertainty in AI

A central challenge for uncertainty communication is that VLMs do not automatically translate their internal uncertainty into language users can act on. To understand why this gap exists, it is helpful to first trace how uncertainty quantification has developed and where it currently falls short for language-visual tasks.

Uncertainty quantification (UQ) is a well-developed area of machine learning concerned with characterizing how much confidence a model's output should carry (Abdar et al., 2021). Foundational work distinguishes *aleatoric* uncertainty, arising from irreducible noise in the data, from *epistemic* uncertainty, arising from gaps in the model's knowledge that could in principle be reduced with more data (Abdar et al., 2021). Classical approaches to estimating both include Bayesian neural networks, Monte Carlo dropout, and ensemble methods;

with more recent work examining on how degrees of belief and disbelief can be expressed directly from evidence (Zhang et al., 2023). These techniques were developed primarily for classification settings, where the output space is discrete and calibration, whether a model's internal confidence scores (token probabilities) align with its actual accuracy, is relatively well-defined.

Captioning tasks introduce a harder problem. When the output is an open-ended natural language description rather than a class label, there is no obvious way to collapse model uncertainty into a single number. Hama et al. address an adjacent version of this problem in image-caption retrieval, showing that Bayesian uncertainty estimates over visual and textual embeddings can identify unreliable retrievals and flag them for abstention (Hama et al., 2021). Their work illustrates both the promise and the limits of internal UQ for language-visual tasks. Particularly how uncertainty can be measured in the embedding space, but the resulting signals are not surfaced to end users and are not interpretable as natural language.

For large language models, internal uncertainty signals — token likelihoods and log probabilities — are rarely exposed directly to users, and are difficult for non-expert users to interpret even when they are (Xiong et al., 2023; Zhou et al., 2023). Xiong et al. evaluated a range of confidence expression methods for LLMs and found that models remain overconfident even when explicitly prompted to verbalize their uncertainty (Xiong et al., 2023). The problem extends to Vision-Language Models: Groot and Valdenegro-Toro show that both LLMs and VLMs exhibit high calibration error and are overconfident most of the time, even after calibration interventions (Groot and Valdenegro-Toro, 2024).

Closer approaches like Mielke et al. introduced the concept of *linguistic calibration*: training conversational agents to emit hedging language ("I'm not sure," "probably," "approximately") in proportion to their actual likelihood of being correct, rather than defaulting to confident-sounding output (Mielke et al., 2022). Their results show that conversational agents are overconfident without explicit calibration, and that training to produce hedges

reduces overconfidence on the tasks they study. Importantly, this work focuses on dialogue systems and fact-based question answering, where outputs can be evaluated against discrete correct answers.

Internal uncertainty signals and externally produced linguistic uncertainty do not track each other reliably, and end users see only the latter. Whether the linguistic surface of a VLM caption, specifically, hedging language, carries meaningful signal about caption reliability is an open question this thesis investigates.

2.2 Explainable AI and Decision Making

Understanding how users respond to linguistic uncertainty in AI output is the second building block for this thesis. If uncertainty is present in VLM captions, its value depends on how users interpret and act on it.

In parallel to the machine learning literature, HCI and explainable AI (XAI) research has examined how explanations shape users' ability to appropriately calibrate reliance on model outputs (Xu et al., 2019; Lage et al., 2019; Ribeiro et al., 2016; Doshi-Velez and Kim, 2017). I draw selectively on this literature, focusing on studies that examine how users interpret *linguistic* uncertainty rather than numerical confidence scores or visual explanation methods. Prior work argues that interpretability mechanisms are necessary for supporting decision-making under uncertainty, as users often struggle to assess when model predictions should be trusted (Doshi-Velez and Kim, 2017). Early work in XAI introduced rule-based explanation methods aimed at improving transparency of black-box models (Ribeiro et al., 2016), while subsequent research has shifted toward evaluating how such explanations affect human decision behavior in practice (Lage et al., 2019). However, much of this literature assumes that uncertainty can be effectively communicated through structured explanations, and pays comparatively less attention to how uncertainty is conveyed through the language itself. In

particular, these approaches leave open how users interpret more subtle cues such as hedging, modality, and epistemic stance in natural language outputs. This thesis therefore shifts focus from explicit explanation formats to understanding the role of linguistic uncertainty in shaping user trust and reliance.

This gap between structured explanation research and linguistic uncertainty is filled by a related line of research on AI-assisted decision making that consistently finds that users interpret linguistic uncertainty cues as signals about model reliability and impacts trust. Kim et al. studied how first-person linguistic uncertainty affect user reliance on LLM outputs (Kim et al., 2024). In their study, prompts framed as *"I'm not sure, but..."* reduced participants' agreement with model answers and improved overall task accuracy. Participants treated the uncertainty as a signal to doubt model output and conduct their own independent verification of the information. Impersonal hedges like *"It's unclear, but..."* were perceived as weaker signals and had smaller effects on reliance. The finding suggests that even subtle grammatical framing of uncertain language shapes how users calibrate trust, beyond presence alone. Extending this line of work to a global context, Rathi et al. examine overconfidence and overreliance across five languages — English, French, German, Japanese, and Mandarin (Rathi et al., 2025). They find that LLMs are overconfident across all languages tested, frequently producing confident language ("it is definitely," "certainly") even in incorrect responses, and that users overrely on confident generations in all five languages. When models do exhibit uncertain language, users tend to reduce their reliance.

Steyvers et al. introduce two concepts that are complementary to the suspected challenges BLV users face when reading VLM captions (Steyvers et al., 2025). The *calibration gap* refers to the difference between a user's confidence in an LLM-generated answer and the model's actual likelihood of being correct. Users consistently overestimate LLM accuracy when interacting with default, confident explanations. The *discrimination gap* refers to the difference between how well the model can distinguish correct from incorrect answers

internally and how well users can make that same distinction from the model's output. Even when a model's internal probability is accurate, those signals are not visible to users who must rely on the language of the response alone. The result is that users cannot reliably tell when the model is likely wrong, even when the model itself could, in principle, signal this. When explanations were modified to better reflect the model's internal confidence (researchers introduced hedges and uncertainty markers where internal probability was low) both the calibration and discrimination gaps narrowed, and users were significantly better at distinguishing correct from incorrect responses. What this means for the current study is that aligning a model's linguistic output to its internal confidence measurably improves user decision-making. What is currently unknown is whether the hedging language VLMs naturally produce in open-ended product captions already achieves this alignment in practice.

2.3 Image Quality and Uncertainty Perception for BLV People

These calibration and discrimination gaps in linguistic uncertainty are particularly consequential in assistive technology contexts, where degraded image quality is common and model errors are correspondingly more frequent. VLMs consistently underperform on low-quality images, largely because most benchmark datasets exclude or under represent low-quality image conditions (Chiu et al., 2020; Dognin et al., 2022). BLV individuals regularly encounter difficulty providing "high quality" images to VLM-based systems as images submitted in real-world conditions often exhibit blur, poor framing, incorrect orientation, partial occlusion, or low lighting (Huh et al., 2024; Karamolegkou et al., 2024).

Alharbi et al. conducted a study of 26 blind people's verification experiences with AI visual assistance technologies, finding that many errors are difficult to detect non-visually and that existing XAI features, which are largely designed for sighted users, provide little support for

blind people trying to assess output reliability (Alharbi et al., 2024). It was found that BLV users do not passively accept or reject AI output but instead, build familiarity with models through repeated use in low-stakes, easily-verifiable scenarios before relying on them for higher stakes tasks, treating these interactions as a kind of calibration process. When they suspect an error, they cross-reference with other devices, employ non-visual sense making (touch, sound), and selectively involve sighted people as a verification fallback (Alharbi et al., 2024).

Furthermore, Adnin and Das studied how blind users of ChatGPT and Be My AI understand and navigate GenAI limitations, finding that users frequently develop flawed mental models of how the tools work and overestimate their reliability in ways that may amplify existing accessibility harms (Adnin and Das, 2024). Crucially, their participants reported reason through competing factors — context, stakes, verifiability, and believability — when deciding how much to interrogate a given output, suggesting that trust is actively negotiated rather than given. Tang et al. frame the resulting experience as "everyday uncertainty", a pervasive mindset of skepticism and criticality towards both AI- and human-mediated assistance (Tang et al., 2025). Their participants did not expect information to be correct or complete; instead, they extracted cues from error-prone sources, treated all information as tentative, and constantly adjusted their strategies based on the politics of access.

BLV users also bring specific, granular information needs to these tools. Jung et al. contribute the idea that BLV people's expectations for image descriptions extend well beyond object identification to include context, brand identity, and details that bear on safety and decision-making (Jung et al., 2022). In the context of product captioning a caption that identifies "a can" is far less useful than one that identifies "a 14.5 oz can of Hunt's diced tomatoes, no salt added." And a hedged caption that says "a can, possibly of tomatoes" may be more useful than either — or may create false confidence if the hedge does not actually reflect the model's accuracy on that image.

Prior work on BLV use of AI visual assistants documents what goes wrong and how users adapt, but it does not examine how the linguistic output of models — specifically, whether models hedge when they are likely wrong. Given that linguistic uncertainty affects trust in sighted users, and that BLV users have fewer fallback verification paths, this gap merits further study.

2.4 Expressing Uncertainty Through Language: Hedging

The previous sections establish that VLMs fail to translate internal uncertainty into reliable linguistic signals, that linguistic uncertainty shapes user trust and reliance, and that BLV users are especially exposed to the consequences of miscalibrated uncertainty language. The question that remains is: what form of uncertainty expression is most likely to be interpretable and useful to BLV users? This thesis focuses on hedging — a form of uncertainty expression pervasive in everyday human communication and produced by language models trained on human text.

Hedging is how speakers signal that they are not fully committed to a claim. Rather than stating something outright, a speaker might soften it — "I think this is the rice," "probably still good," "roughly 200 calories" — to signal that their knowledge is partial or their confidence is limited (Lakoff, 1973). Corpus linguistics research shows that hedging is a routine feature of human speech across registers. Common hedging devices in American English discourse appear at rates ranging from 5% to 27% depending on the device and context, making hedging one of the most frequent communicative strategies in everyday language (Prokofieva and Hirschberg, 2014).

Prokofieva and Hirschberg distinguish two ways this softening works (Prokofieva and Hirschberg, 2014). *Relational hedges* or "shields" distance the speaker from a claim by attributing it or

framing it as personal view (*I think*, *it seems to me*). *Propositional hedges* or "approximators" fuzz the content itself (*sort of*, *about*, *roughly*). A speaker can distance themselves from the claim as a whole — "I think," "it seems to me" — which signals uncertainty about whether the claim is right. Or they can fuzz a specific part of the content — "about," "roughly," "possibly" — which signals that a particular detail is imprecise. This is particularly valuable in the context of product captioning because VLMs do both: a caption like "this appears to be a can of soup, possibly chicken noodle" combines uncertainty about the product identity with uncertainty about the variety, and those two uncertainties convey different signals for a BLV user deciding whether to act on the caption.

To prior knowledge, no work in linguistics or NLP examines hedging in relation to blindness, visual impairment, or accessibility contexts. Linguistic literature treats hedges as a general feature of human communication while computational work treats hedge detection as an information-extraction problem. Neither asks what it means for a user who cannot independently verify a claim when the model hedges on it — or fails to hedge when it should.

Pulling from the conceptual work in each of these disciplines, I next aim to examine the extent that hedging occurs within VLMs for product captioning tasks and the role this plays in how BLV users interact with models in this context.

3 Methods

This study analyzes a dataset of real-world product images to examine how VLMs describe products and how hedging relates to caption accuracy and image quality. We begin with the VizWiz dataset, filtering images to those where products are likely identifiable and further refining the dataset through manual review to ensure clear, verifiable product presence. The resulting dataset is split into high and low-quality image subsets and annotated with structured product information. We then generate captions using four VLMs with identical hyperparameters and prompts. All captions were manually evaluated for product identification accuracy, and a subset is additionally annotated for hedging language using a validated detection pipeline. Finally, we apply descriptive statistics and logistic regression models to examine relationships among hedging, accuracy, image quality, and model identity.

3.1 Dataset

This section is adapted from prior work published at CHI26 that I am a co-author on (Garg et al., 2026).

Data Selection

To create a dataset focused on products, we start with Gurari et al.'s VizWiz Image Captioning dataset (Gurari et al., 2020), which includes five crowdworker-provided image captions on photos taken by blind people, and Chiu et al.'s VizWiz Image Quality Assessment

dataset (Chiu et al., 2020), which includes annotations on image quality issues by five crowdworkers. Using their training dataset (23,431 images), we first filtered based on whether humans can reliably identify the pictured product, indicating that image quality issues are not severe enough to prevent image description. We selected these data by including images where two or fewer crowdworkers indicated that the image was unrecognizable (conversely, three to five crowdworkers provided a caption). Upon inspecting the dataset, we noticed that most images of products included text; therefore, we only included images where crowdworkers identified text in the image as a heuristic for product identification. This resulted in a filtered dataset of 14,398 images (61.5% of the original dataset).

We then created two subsets of data. First, we focused on *high-quality* images of products without quality issues. This dataset serves as a benchmark for evaluating the performance of VLMs in product identification on natural images. We selected images for which 4 or 5 crowdworkers flagged the image as having no issues, and at most 1 person flagged each image quality issue, resulting in 2,599 images (11.1% of the original). Second, we created a dataset of *low-quality images*, where 4 or 5 crowdworkers flagged the image having an *image quality* issue (blur, rotation, framing, obstruction, or being too bright or too dark), resulting in 5,432 images (23.2% of the original). This dataset corresponds to images for which human captioners felt confident providing a caption, despite identifying image quality issues that could potentially hinder their accuracy.

To understand how accurately VLMs identify products, we manually reviewed all images to identify those that appeared to contain products. Four researchers reviewed all images in each subset. They excluded images that did not include products, such as nondescript boxes or pictures of rooms in the home. We excluded images of computer screenshots, currency, printed papers, books, CDs, DVDs, clothing, and unpackaged electronic devices. These were, on the whole, difficult for annotators to objectively verify, such as identifying an

article of clothing or the name of a book based on a page of its text.[1] We also excluded any images where more than one product is pictured. For the high-quality images specifically, we also eliminated any product images with even mild distortions (e.g., camera blur; lens flares) to ensure the subset was completely free from image quality issues. This resulted in a high-quality subset of 729 images and a low-quality subset of 1,696 images.

Data Annotation

The survey results showed that BLV people want specific product information when selecting foods, medicines, and personal products. We developed a three-part annotation scheme consisting of:

- Product: the generic term for the product (e.g., cereal, soup, meal, medication).
- Brand: any detectable brand information (e.g., Betty Crocker, Kraft, Great Value, Kellogg's).
- Variety: details about the type, flavor, or variety (e.g., containing peanuts, low sodium)

A team of four researchers manually annotated each image using this structure. When annotating a product, researchers reviewed the image and human captions. If unsure, researchers also searched online for product images or noted that they were unsure about the image, so another researcher could review it. The image was excluded from the dataset if researchers were uncertain about the pictured product. For example, we excluded images that showed only product barcodes or lacked the visible details required for product verification. To ensure the validity of product annotations, a second researcher then reviewed and confirmed agreement with each image and annotation. Any discrepancies were flagged for discussion, and if no agreement was reached, the image was removed from the dataset. To enable con-

[1]While these are real-world cases where objects are ambiguous and valuable to identify, we require images with clear, correct, and assessable annotations to understand how VLMs fail (our focus), where more ambiguous or hard to verify examples could create a confound in our analysis.

sistency in product naming, the research team aimed for the most specific name within the product (e.g., granola instead of cereal, Sprite instead of soda) and included both brand and sub-brand names when available. When possible, we included known flavor or ingredient details (e.g., vanilla soymilk; chicken with potatoes and green beans) and details related to dietary needs or potential allergies (e.g., zero-sugar Gatorade, peanut butter granola bars). This detailed, validated labeling is distinct from coarse object and product labels in prior work (Gurari et al., 2018; Kacorri et al., 2017). This process yielded a final dataset of 1,859 images annotated with product details, comprising 729 high-quality and 1,130 low-quality images.

Generating Captions From VLMs

We used four different VLMs to generate captions for our dataset. We include GPT-4.1 since the three most commonly used AI tools in our survey—Seeing AI (Beatman and Leen, 2024), Be My AI (Be My Eyes, 2023), and OpenAI's ChatGPT (OpenAI, 2025a)—all use a GPT-4 class model from OpenAI. We include Google's Gemini 2.5, another frequently used model. Finally, we include two recently released open-source models: Llama from Meta (Grattafiori et al., 2024)[2] and Molmo from the Allen Institute for AI (Deitke et al., 2025), which both exhibit comparable performance to closed-source industry models on benchmarks. We include open-source models because BLV people in our survey and prior work (Stangl et al., 2022, 2023) expressed concerns about data privacy when using LLMs—which open-source models can address when run locally. Moreover, open-source models provide access to the model architecture, training regime, and, in some cases (e.g., Molmo), training data, affording greater flexibility for improving performance than closed-source models. For GPT-4.1, we used OpenAI's API and selected the gpt-4.1-2025-04-14 model checkpoint for reproducibility (OpenAI, 2025b). For Gemini, we used Google's API for gemini-2.5 (Google, 2025); Google does not provide a more specific model checkpoint. For Llama and Molmo, we used Llama-3.290B-

[2]The Ray-Ban Meta Glasses, the most used general-purpose AI tool after ChatGPT, are also powered by a version of Llama (Meta, 2025)

Vision-Instruct (Meta, 2024) and Molmo-72B-0924 (AllenAI, 2024) from Hugging Face, with 4-bit quantization.[3] Llama and Molmo were run locally on two NVIDIA RTX A6000 GPUs. For all models, we set `temperature = 1.0` and `top_p = 0.95` to balance determinism and randomness of output generation[4], and `max_new_tokens = 500` for generated tokens to allow for detailed captions. Before generating captions, images were converted to PNG with the alpha channel removed—since some VLMs can perform poorly with transparent images—but no additional processing was done (e.g., blur reduction; image super-resolution). For brevity, we refer to these models as "GPT", "Gemini", "Llama", and "Molmo" throughout the rest of the paper.

We instructed each VLM to caption each image with the same prompt (Garg et al., 2026). Our prompt was inspired by prior work using VLMs to describe images for BLV people (Mohanbabu and Pavel, 2024; Chang et al., 2024; Huh et al., 2024), and developed following best practices (OpenAI, 2025c). We focused on prompting the VLM to identify key features, such as the object, product type, brand names, and variety details essential to understanding the product in the image while abstaining from vague language.

Dataset Coding

As the final step in our dataset creation process, we manually verified the correctness of each VLM-generated caption. We performed human coding due to the issues with existing captioning metrics and because LLM-as-judges—while correlating well with human judgment for simple question-answer tasks (Zheng et al., 2023)—may be falsely lenient on more open-ended tasks, like slightly incorrect product descriptions (e.g., Coke Zero versus Diet

[3]We tested the smaller Llama-3.2-11B-Vision-Instruct and Molmo-7B-O-0924 with 16-bit precision, but found that the larger, quantized models performed better while fitting within our compute limitations. Prior work suggests that performance loss is marginal with 4-bit quantization (Dettmers and Zettlemoyer, 2023; Frantar et al., 2023).

[4]We tested various `temperature` (0 - 1.0) and `top_p` (0 - 1.0) settings. `temperature` had little effect on product identification quality. In contrast, `top_p` caused more noisy captions above 0.95 (the default for our models). These settings are similar to prior work that has used VLMs for captioning (Chan et al., 2023; Nguyen et al., 2023).

Coke) (Thakur et al., 2025). All VLM captions were anonymized to minimize potential bias during coding (i.e., Models A, B, C, D), and any image metadata (e.g., what quality issues were present) was concealed, except for the product annotations. The order of images was also randomized to reduce any ordering effects.

Four researchers coded the accuracy of the four VLM captions for each of the 1,859 images in our dataset (7,436 captions in total). Before coding, the research team met and developed a coding scheme that allowed for minor spelling mistakes and term variation (e.g., soda vs. soft drink; chips vs. crisps) but was strict on key details (e.g., brand and sub-brand; ingredients when describing food variety). Captions were marked as incorrect if there were major hallucinations (e.g., 12 ounces reported as 12-pack, for a soda can) or contained errors that changed their meaning (e.g., grilled chicken instead of fried chicken). Each researcher coded a sample of 50 randomly selected images with all VLM captions (a total of 200 captions). IRR was computed using Krippendorff's alpha, with an agreement of 0.859. Following the training period, the four researchers independently coded the remaining images, marking ones they were unsure about as "maybe". The team reviewed and discussed these and other challenging cases.

Hedge Detection and Annotation

As defined in the Background, hedging is how speakers signal that they are not fully committed to a claim (Prokofieva and Hirschberg, 2014). Examples include epistemic modals (might, could, may), tentative verbs (appears, seems, suggests), and approximators (about, roughly, approximately).

To identify hedging automatically, a regular expression based pipeline was designed to capture a broad range of hedge cues (see Table 3.1).

Table 3.1: Hedging Annotation Procedure

Stage	Description
Initial Pass	A wide-coverage regex list was constructed using hedging markers from computational linguistics literature (see Appendix A).
1st Review	Captions were manually reviewed to identify false positives, a caption that contains a hedge matched by the regex but where the matched term does not function as a hedge in context. For example, the word *about*, a common approximator, would be flagged by a hedge regex but in a caption like "information about the product" it is a preposition, not a hedge.
Refinement	Regex expressions were revised twice to improve precision, yielding a final annotation where `model_hedge` $= 1$ for any caption containing one or more validated hedge markers.
2nd Review	A second round of manual review was conducted to correct remaining false positives.

The initial regex list was constructed to maximize recall, capturing all plausible hedges. For example, patterns matching *likely*, *appears to be*, *possibly*, *approximately*, and *I think* across varied contexts. Manual review then identified and removed false positives introduced by the broad initial pass, and expressions were revised accordingly. A second manual review corrected any remaining errors.

To ensure that annotations were applied consistently across researchers, interrater reliability was assessed on a shared subset of captions containing hedging language. Two researchers

coded a shared sample of 100 randomly selected captions that contained hedging (200 captions total) to establish reliability. Remaining images were coded independently, with uncertain cases flagged and discussed as a team. Reliability thresholds of Cohen's $K > 0.70$ were used to indicate acceptable agreement (Cohen, 1960).

3.2 Measures + Analysis

The analysis examines how hedging relates to caption accuracy and image quality across models, using a combination of descriptive statistics and logistic regression. All outcomes are binary: hedge presence (0/1) and caption accuracy (0/1), making logistic regression the appropriate modeling choice.

Descriptive Analysis

Baseline hedge rates were computed per model and separately for correct and incorrect captions to characterize cross-model variation and the direction of the hedging-accuracy relationship prior to regression.

Inferential Analysis

Three regression models were estimated: two single-predictor models evaluating hedging and image quality as independent predictors of caption accuracy, and one unified model examining how hedging behavior is jointly shaped by accuracy, image quality, and model identity. The Akaike Information Criterion (AIC) metric was used during model development to compare each candidate model and determine final model parameterizations (Akaike, 1998).

Model 1 — Caption accuracy predicted by hedging. A single-predictor model with caption accuracy as the outcome and hedging as the sole predictor, estimated on all captions pooled across models. This model characterizes the association between hedging and

correctness by model identity.

Model 2 — Caption accuracy predicted by image quality. A single-predictor model with caption accuracy as the outcome and overall image quality (high vs. low) as the sole predictor, estimated on all captions pooled across models. This model establishes the baseline of image quality as a predictor of caption error, providing the comparison against which hedging (Model 1) is evaluated.

Model 3 — Hedging predicted by image quality, accuracy, and model identity (Unified Model). This is the primary model. The outcome is hedge presence; the dataset is in long format with one row per caption (7,436 observations: 1,859 images × 4 models). Predictors are:

- **Model identity**: three variables (Gemini, Llama, Molmo), with GPT-4.1 as the reference category. GPT-4.1 serves as the reference because it exhibits the strongest alignment between hedging and accuracy in descriptive analysis, providing a baseline against which the other models' deviations can be assessed.
- **Image quality labels**: the eight binary indicators listed above, coded at the image level.
- **Caption accuracy**: binary (0/1), coded at the caption level and varying across models for the same image.
- **Interaction terms**: accuracy × model identity (three terms) and image quality × model identity (retained based on AIC fit). These interactions allow the relationship between hedging, accuracy, and quality to vary by model.

For models containing interaction terms, interaction plots were constructed to support interpretation. Each plot displays the predicted probability of hedging across levels of the

focal predictor (e.g., blur, framing, accuracy), separately for each model, with confidence intervals.

Prior to interpreting regression results, the independence of observations was assessed at the caption level; the within-image correlation introduced by the long-format structure (four rows per image) is addressed separately as described above. Multicollinearity among image quality predictors was assessed using variance inflation factors (VIF); all eight quality issue labels had VIF < 1.21, indicating no collinearity among image quality predictors. Model identity dummies and the accuracy term showed elevated VIF (range 4.1–5.8) due to the inclusion of accuracy×model interaction terms.

For all models, final parameters were selected using sequential AIC refinement. A set of predictors was retained only if it reduced AIC by at least 2 points relative to the previously accepted model; the null model (intercept only) was used as the baseline for per-VLM models. Model fit statistics for all selected specifications are reported alongside coefficient estimates in the Findings section.

It is important to note that hedging and accuracy are both properties of the same caption output and cannot be interpreted in a causal way within this design. Both may reflect a shared upstream condition — the model's processing of a difficult image — rather than one driving the other.

Per-VLM Models and Model Selection

A separate logistic regression was estimated for each model with hedge presence as the outcome and image quality labels as predictors. This per-VLM specification allows more characterization of which quality conditions predict hedging within each model, without the cross-model interactions required by the unified model. Parameter selection followed the same AIC-based procedure described above, with the null model (intercept only) serving as the baseline for each per-VLM specification.

4 Findings

A central premise of this work is that users do not encounter a model's internal uncertainty directly but instead, they encounter language in the output. This analysis answers whether a model's natural language expressions of uncertainty correspond to accuracy and image quality issues.

4.1 RQ1: Hedging Occurs Across Models but Varies in Use

Hedging is present across all four models but varies in frequency. Figure 4.1 shows the overall hedging frequency for each model. Across models, Llama hedges in 17.4%, Molmo in 11.9%, GPT in 6.3%, and Gemini in only 0.8%. If users treat hedging as a proxy for confidence, then the interpretive landscape varies substantially depending on which model they interact with.

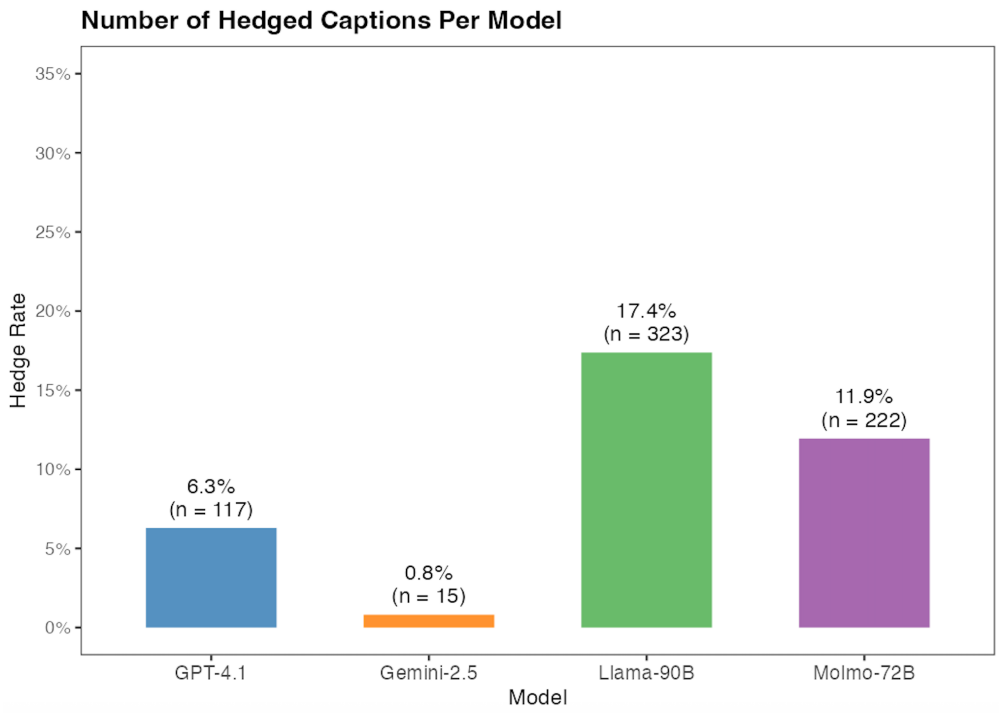

Figure 4.1: Variation in hedging frequency across VLMs. Llama hedges in 17.4% of captions, Molmo in 11.9%, GPT in 6.3%, and Gemini in 0.8%.

Differences in baseline hedging frequency are likewise mirrored in the choice of hedge words (See Figure 4.2). GPT exhibits the most consistent use of epistemic hedges (e.g., "likely" and "appears to be"), whereas Llama shows broader spread, including likelihood and approximation terms (e.g., "around," and "approximately"). Gemini uses fewer hedge words overall, with a concentration in approximation terms.

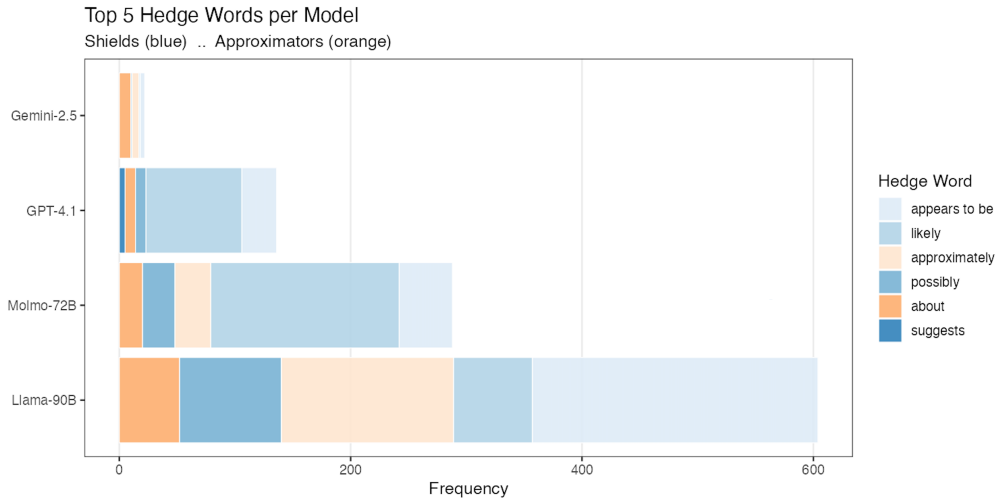

Figure 4.2: Top hedge words used by each model, grouped into two categories: approximators (orange; e.g., "approximately," "about") and shields (blue; e.g., "likely," "appears to be," "possibly," "suggests"). GPT primarily uses epistemic hedges such as "likely" and "appears to be," indicating uncertainty about correctness. Llama shows a broader distribution across both categories, using both likelihood and approximation terms. Gemini uses fewer hedge words overall, with most of its hedging concentrated in approximators.

Hedges convey varying *kinds* of uncertainty. Prior work shows that hedging functions as a way for speakers to adjust their commitment to a claim through specific word choice (Lakoff, 1973; Prokofieva and Hirschberg, 2014). For instance, approximators such as *about* or *approximately* modify the category membership or precision of a description. In a captioning context, a model might say *"approximately 5 ounces per serving"*, signaling to a BLV reader that the object quantity is loosely specified. In contrast, modal expressions such as *might* or *likely* indicate uncertainty about the truth of the statement itself. A caption such as *"a can, likely diet coke, is placed on a counter"* communicates uncertainty whether the interpretation of the image is correct. When models rely on different hedge vocabularies despite encountering similar visual ambiguity, BLV users may interpret their confidence differently even when the underlying accuracy is comparable. This allows the possibility that hedging language could lead to inconsistent or misleading interpretations.

The following examples illustrate how these vocabulary differences play out in practice —

one case where the hedge type matches the visual uncertainty in the image, and one where it does not.

Image / Model	Caption
	**GPT:** Plastic cup containing 6 ounces (170 grams) of fat-free yogurt. The packaging is mostly white with a bright pink section and a blue fruit graphic, **likely indicating a fruit flavor such as blueberry.** The label mentions "2X the Calcium of the leading yogurt" and "50% Daily Value in every cup." **The brand and exact flavor are partially obscured.** **Product:** Yogurt **Brand:** Yoplait **Variety:** Blueberry
	**Llama:** The product is a can of Coca-Cola, featuring a red color with white lettering and a distinctive shape. The can has a diameter of **approximately** 3 inches and a height of 4.5 inches. **Product:** Crackers **Brand:** Cheez-it **Variety:** N/A

Table 4.1: Qualitative examples illustrating aligned and misaligned hedging in VLM-generated captions. **Row 1 (top):** An image of a Yoplait yogurt cup (partially framed, with brand and flavor text obscured). GPT-4.1 correctly identifies the product as yogurt and uses hedging appropriately on the ambiguous attribute ("likely indicating a fruit flavor such as blueberry"), aligning uncertainty with what is visually unclear. **Row 2 (bottom):** An image of a Cheez-It cracker box (poorly framed and slightly blurred, showing the back of the package). Llama-90B incorrectly identifies the product as a Coca-Cola can and applies hedging to an irrelevant attribute ("approximately 3 inches"), misplacing uncertainty on a hallucinated detail.

In Table 4.1, the GPT caption illustrates hedging that is appropriately matched to the image. The yogurt cup is correctly identified, but the brand and variety are cut off due to framing. GPT hedges precisely on the attribute the image leaves ambiguous, "likely indicating a fruit flavor such as blueberry", while stating the product type and nutritional claims confidently. Notably, the caption is correct despite the hedging in this example.

The Llama caption illustrates the inverse failure. The image shows the back of a Cheez-It box, poorly framed and slightly blurred. Llama misidentifies the product entirely, asserting it is a can of Coca-Cola, while hedging on the can's diameter ("approximately 3 inches"). The approximator is placed on a physical measurement of a hallucinated product. This pattern appeared consistently across Llama captions: approximators applied to peripheral measurements or dimensions rather than to the product identification itself. In this case, the pattern is particularly problematic given the placement of the hedging language in no way indicates what is actually correct in the image.

Given these differences in baseline frequency and word usage, we next ask whether these hedging patterns *meaningfully* reflect model accuracy.

4.2 RQ2: Hedging Co-occurs with Inaccuracy but Unreliably

Hedging appears more among incorrect captions for most models. Table 4.2 shows hedge rates by model and caption accuracy. GPT shows the sharpest contrast: 19.0% of incorrect captions contain a hedge versus 3.9% of correct ones. Llama and Molmo show weaker but consistent patterns (Llama: 21.7% vs. 14.7%; Molmo: 15.5% vs. 9.1%). Gemini shows virtually no difference (0.9% vs. 0.8%).

Table 4.2: Hedging frequency by model and caption accuracy, showing the proportion of captions that contain hedging language for both correct and incorrect outputs. GPT-4.1 exhibits the largest difference, with hedging in 19.0% of incorrect captions (56/295) compared to 3.9% of correct captions (61/1,564). Llama-90B (21.7% vs. 14.7%) and Molmo-72B (15.5% vs. 9.1%) . Gemini shows virtually no difference, (0.8%, 12/1,507) and incorrect captions (0.9%, 3/352).

Model	**Correct Captions**	**With Hedge (%)**	**Incorrect Captions**	**With Hedge (%)**
Llama	1,126	165 (14.7%)	733	159 (21.7%)
Molmo	1,041	95 (9.1%)	818	127 (15.5%)
GPT	1,564	61 (3.9%)	295	56 (19.0%)
Gemini	1,507	12 (0.8%)	352	3 (0.9%)

What this means is that, for GPT, when it uses hedging language, the caption is substantially more likely to be wrong. In contrast, for Llama and Molmo, the hedge-accuracy relationship is weaker as both models hedge at similar rates across both correct and incorrect captions. Gemini hedges so rarely that we are unable to interpret.

The per-model rates show that hedging and error tend to co-occur, but they do not say how useful that co-occurrence actually is. If a user saw a hedged caption, how often would they be right to suspect an error? A logistic regression (See Table 4.3) indicates that hedged captions are less likely to be correct ($\beta = -2.41$). Given an AUC of 0.61, there is limited evidence to distinguish the effect from other factors. While hedging is statistically *associated* with error, it does not *reliably* identify error. Many incorrect captions contain no hedge, and many hedged captions are correct. If a user interprets hedging as an indication of accuracy, this poses a concrete risk. A BLV user encountering two captions for the same product — "likely a soup product" and "a chicken noodle soup product" — has no reliable basis for distinguishing them. One contains a hedge while the other does not, but both are similarly likely to be incorrect. Given that confident-sounding language recruits greater user trust regardless of accuracy (Steyvers et al., 2025; Kim et al., 2024), a BLV user interacting with

the unhedged caption has no linguistic reason to question it.

Table 4.3: Logistic Regression Model Results

Model	**Coefficient**	**Intercept**	**AUC**
Accuracy ∼ Hedge	-2.4084	1.8108	0.6138
Accuracy ∼ Quality	-1.0311	2.8482	0.7856

To further capture this relationship as a single comparable number, **hedging accuracy** is defined as the proportion of captions where hedging presence and correctness agree (hedged-and-wrong or unhedged-and-right). GPT scores 97.1%. Gemini scores 95.1%, but this number is largely uninterpretable given that most Gemini captions are unhedged-and-correct by default, not because hedging is well-calibrated. Llama (77.4%) and Molmo (81.6%) score lower, reinforcing what previous rates already suggest: these models hedge frequently enough that many correct captions still contain a hedge.

Model	**Hedging Accuracy (%)**
GPT	97.1%
Gemini	95.1%
Llama	77.4%
Molmo	81.6%

Table 4.4: Baseline hedging accuracy across VLMs for high-quality images.

The model-specific patterns above suggest that the accuracy-hedging relationship, in addition to overall frequency and hedge choice, is likewise not uniform across models. To examine whether this variation holds after accounting for image quality and model identity simultaneously, a **unified** logistic regression was constructed (See Table 4.5), predicting hedge presence from image quality issues, accuracy, and model identity. This model includes interactions with both accuracy and model. GPT was chosen to serve as the baseline category due to its high hedging accuracy.

The coefficient for caption accuracy is negative and significant ($\beta = -1.53$, $p = .000$),

reinforcing the claim that, for the reference model (GPT), correct captions are less likely to contain hedging, accounting for image quality and model identity simultaneously. What is also reinforced is that relative to GPT, Gemini hedges far less ($\beta = -2.4$, p= .003) and Llama hedges far more ($\beta = 1.01$) and that simply the model type remains more impactful as indicators of hedging language.

Variable	**Coef**
Main Effects	
Intercept	-2.51***
framing	0.30
blur	1.06***
rotation	0.16
accuracy	-1.53***
model_gemini	-2.46**
model_Llama	1.01**
model_molmo	0.43

Table 4.5: Main effects from the unified logistic regression predicting hedging behavior. Significance levels: *** $p < .001$, ** $p < .01$, * $p < .05$.

When looking at the interaction effects in Table 4.6, we see variation in this relationship by model. The accuracy $\times$ model interactions indicate that this effect is significantly lower for Llama and Molmo (Llama: $\beta = 1.06$, $p < .001$; Molmo: $\beta = 1.01$, $p < .001$), suggesting a weaker relationship between accuracy and hedging in these models. In contrast, the accuracy $\times$ Gemini interaction does not reach significance ($\beta = 1.42$, $p = .056$), consistent with Gemini's overall low hedge rate across conditions.

Variable	Coef
Interaction Effects	
framing:accuracy	0.63***
blur:accuracy	-0.15
rotation:accuracy	-0.25
framing:model_gemini	-0.69
blur:model_gemini	-2.21**
rotation:model_gemini	1.10†
framing:model_Llama	-0.61*
blur:model_Llama	-0.77**
rotation:model_Llama	0.35
framing:model_molmo	-0.32
blur:model_molmo	-0.47†
rotation:model_molmo	0.03
accuracy:model_gemini	1.42†
accuracy:model_Llama	1.06***
accuracy:model_molmo	1.01***
Model Fit Statistics	
Sample Size = 7,436 observations (1,859 images × 4 models)	
Null Deviance = 4539.8, Residual Deviance = 3910.1, AIC = 3986.1	

Table 4.6: Interaction effects from the unified logistic regression predicting hedging. NA indicates coefficients not estimated due to singularities. Significance levels: *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$.

Molmo: "A white rectangular box, **likely a cigarette pack**, held in someone's hand. The box has "Made in USA" in gold text and a prominent red label with "Marlboro" in white letters. It features a sun-like element above the brand name and "menthol" in gold text below."

Figure 4.3: Molmo caption for a Cheyenne cigarette pack being held in a hand, showing a red-and-white branded box.

Figure 4.3 depicts a Cheyenne brand cigarette pack. Molmo identifies the product category correctly ("likely a cigarette pack") but asserts the brand as Marlboro with no qualification, despite the brand being completely wrong. Hedging language is present, but it is on the one claim that happens to be right. The hallucinated brand, the information a BLV user would likely seek to verify, carries no uncertainty signal at all. This illustrates what the rates in Table 4.2 suggest of incorrect captions frequently appearing no different from correct ones.

This section suggests that hedging alone is a weak predictor of accuracy (AUC = 0.61, Table 4.3). The next asks whether hedging responds to specific image quality issues rather than accuracy alone.

4.3 RQ3: Image Quality Co-occurs with Hedging Inconsistently

Considered across all models, image quality predicts caption accuracy substantially better than hedging does (AUC = 0.79 vs. 0.61, Table 4.3), and yet specific quality issues still do not reliably co-occur with the hedging that could warn BLV users about errors.

When considering the effects per-VLM, blur most consistently occurs with hedging (See Table 4.7), and this relationship holds for GPT and Molmo but not for Llama or Gemini. The pattern is model-dependent, as for most models, hedging does not correspond to the type of image quality issues.

Quality Issue	**GPT**	**Molmo**
Blur	1.225***	0.610***
Framing	0.771***	0.294†
Rotation	0.338	0.240
Intercept	-4.081***	-2.518***

Table 4.7: Per-VLM logistic regression coefficient estimates (M1: hedge $\sim$ quality). *** $p < .001$, ** $p < .01$, * $p < .05$. Gemini and Llama selected the null model (M0) are excluded from the results.

The framing $\times$ accuracy interaction in the unified model in Table 4.6 ($\beta = 0.63$, $p = .001$), suggests that under framing image quality issues, the hedging rate between correct and incorrect captions decreases. Correct captions hedge at closer to the rate of incorrect ones, reducing the distinction.

Having established which quality conditions co-occur with hedging, I next examined hedging accuracy under these same low-quality conditions. Hedging accuracy lowers under blur but the pattern is uneven across models as seen in Table 4.8. For GPT, hedging accuracy

is 70% under blur and reaches 73% under rotation, indicating that, for GPT, agreement between hedging and correctness weakens under blur. Llama shows consistently poor hedging accuracy under most image quality issues (e.g., 47% under blur, 46% under framing), though this occurs alongside high baseline hedging rates. Molmo similarly shows high hedging accuracy under blur (40%). Gemini's hedging behavior remains largely unresponsive to image quality. Despite clear drops in accuracy under degraded conditions, its hedge rate stays near zero, meaning that Gemini's scores are not directly comparable to models like GPT or Molmo. Overall, these results indicate that some image quality issues correlate with error among hedged outputs, but the effect and reliability vary by model and quality issue.

Quality Issue	**GPT**	**Gemini**	**Llama**	**Molmo**
Framing (n=176)	74%	69%	46%	41%
Blur (n=646)	70%	69%	47%	40%
Rotation (n=430)	73%	69%	51%	32%

Table 4.8: Hedging accuracy under low-quality conditions. Hedging accuracy indicates the proportion of captions where hedging presence and correctness agree (hedged-and-wrong or unhedged-and-right.)

Summary

The findings across the three research questions reveal that hedging in VLM product captions is variable and shaped more by which model generated the caption than by the quality of the image that caption was produced from. Hedging co-occurs with inaccuracy at rates that are statistically meaningful — most clearly for GPT, least clearly for Gemini — but the association is too weak to function as a reliable indicator of error on its own. Image quality predicts caption accuracy more strongly than hedging does, yet the same quality conditions that most reliably produce errors do not most reliably co-occur with uncertainty language. The result is a gap between what makes a caption wrong and what makes it sound uncertain.

5 Discussion

This thesis set out to evaluate how, when, and how reliably VLMs hedge in product-identification captions for BLV users. What was found is that hedging in VLM-generated product captions is not driven by low-quality images, but by model identity. Image quality reliably predicts caption error, but does not reliably produce the hedging that would allow users to anticipate that error. This disconnect suggests that hedging, as currently produced, appears to be more strongly shaped by training data and alignment practices than by input uncertainty. The sections that follow work through what this means: for model design, for how we think about communicating uncertainty in assistive AI, and for BLV users navigating these assistive technologies where uncertainty is a daily condition.

5.1 Reliability of VLM Hedging

Hedging behavior is likely a product of training, not model uncertainty

Model identity was found to be the strongest predictor of hedging, exceeding all image quality variables in the unified regression. The range, from Gemini's near-total absence of hedging to Llama's rate of 17.4%, seems to reflect differences in how model communicative behavior was shaped during training and alignment.

The four models in this study differ in architecture, training data, and fine-tuning procedures, and these differences are the most likely explanation for their divergent hedging behavior.

Molmo, for example, was trained on the PixMo-Cap dataset, which contains human-authored captions with naturally occurring hedging language (Deitke et al., 2025). Models learn to minimize the difference between their output and their training data so if hedges are present in the training captions, they will appear in the output. The issue here is that models have no explicit representation of *why* the hedging was there in the first place, whether the original writer was genuinely uncertain or being polite (Prokofieva and Hirschberg, 2014).

For the closed-source models (GPT and Gemini), training data and fine-tuning details are not publicly available, which limits the inferences that can be drawn about *why* they hedge as they do. What can be observed is the output pattern: GPT hedges selectively and in closer association with error, while Gemini almost never hedges. These differences likely reflect different training procedures, but attributing them to any specific technique would require access to information that is not available for these models.

Hedging is an unreliable proxy for correctness

The association between hedging and inaccuracy (strongest in GPT, weaker in Llama and Molmo, negligible in Gemini) establishes that hedging and accuracy co-vary in ways that are partially consistent. The regression coefficient for accuracy ($\beta = -1.53$, $p < .001$) reflects this co-occurrence controlling for model identity and image quality: correct captions are less likely to hedge.

Huh et al. (Huh et al., 2024) found that BLV evaluators rated incorrect VLM-generated answers as significantly more plausible than sighted evaluators did, a finding attributed to the absence of independent visual verification. Karamolegkou et al. (Karamolegkou et al., 2024) identified confident responses when a question could not be answered as a critical safety failure in assistive applications. Both observations contextualize the importance of the current findings as models that produce confident sounding incorrect captions without hedging provide BLV users with no linguistic basis for doubt.

This likely matters more for BLV users than the current findings alone suggests, as research on how blind people navigate GenAI errors shows that they actively reason about which parts of a response to trust, weighing context, stakes, and prior experience with the model (Adnin and Das, 2024). A hedge positioned on the product category ("possibly a can of soup") carries different weight than a hedge positioned on the variety ("possibly chicken noodle"), even if both come from a caption scored as "incorrect" in an accuracy analysis. Getting the hedge in the right place — on the part of the caption the model is actually uncertain about — is a design requirement that future work should explore further.

Models fail to consistently communicate the image quality conditions that cause their errors

Garg et al. document that image quality degrades VLM accuracy substantially — from 98% on high-quality images to 75% when quality issues are present, with further drops as issues compound (Garg et al., 2026). The present findings supplement these earlier findings, showing that the same quality degradations that most reliably produce errors are the least likely to produce hedging. The gap between what makes a caption wrong and what makes a caption sound uncertain is evidence that hedging, currently, cannot serve as a stand-alone indicator of accuracy for BLV users. In other words, a caption can be wrong in exactly the way one would expect from a low quality image, and still appear no different from a correct one.

Blur is a notable partial exception. As the most common image quality issue in BLV photography, blur elevates hedging in GPT and Molmo, the two models that show meaningful quality-hedging sensitivity. The regressions confirm that the framing, rotation, and obstruction image quality issues do not produce the same effect. For Gemini and Llama, the regressions indicated that quality does not predict their hedging at all. Blur, precisely because it is both common in BLV photography and partially associated with hedging in two models, could serve as a high-priority target within prompting strategies — conditioning the

model's language on detected blur before generation.

5.2 Recommendations for Aligning Hedging with Model Accuracy

The findings above establish what is missing in current VLM captioning behavior and why it matters for BLV users. The remainder of this discussion pulls the design implications out of the findings and organizes them by where in the VLM pipeline the intervention would apply.

Provide model-specific guidance on how to interpret hedging.

Because hedging patterns vary substantially across VLMs, the trust building work BLV users do when they first adopt a new AI tool does not transfer across models. In cases like these where models have the tendency to error, BLV users build familiarity with models through repeated use in low-stakes, using these early interactions as a kind of informal trust building (Alharbi et al., 2024). A BLV user who has learned that GPT's hedges are meaningful signals may be poorly acclimated if they switch to Llama, whose hedges are far less suggesting of accuracy. Given that hedging patterns are model-specific, then this trust building process should likely be model-specific too. Model-specific guidance, like explicit documentation or live communication of a particular model's hedging language alignment, could serve as a potential avenue for model interface design. Furthermore, assistive applications that use multiple VLMs should provide explicit mention of when and precisely *where* current model's hedging language conflicts with internal uncertainty in a caption, rather than leaving users to attempt to discover this through trial and error. This guidance should be understood as scaffolding within existing BLV user practices. Referring back to Tang et al. framing of BLV users' relationship to uncertain information as "everyday uncertainty" (Tang et al., 2025), interface design should support this work rather than replace it by integrating cross-

referencing across models where feasible.

Alongside the user interaction with *existing* model error, we next give recommendations for the *future* how hedging can be aligned at the foundational level.

Treat hedge accuracy as a post-training feature

Given that these models will hedge, how can we make hedging actually useful? At the training level, fine-tuning on examples that pair specific quality issues with appropriately hedged captions could address this issue, producing models whose uncertainty expression responds to input. But training interventions on the model side are only half of the problem. The other half is whether the resulting uncertainty is communicated to users in a way they can use.

More precisely, the wearable assistive system described by Ruan et al. demonstrates that explicit error-flagging and corrective prompting are feasible at the interface level. The present findings establish a rationale for extending that logic into the language of the caption itself (Ruan et al., 2026). Current standard-training benchmarks reward accuracy but do not reward hedging placed on the specific claim most likely to be wrong. Loss functions that evaluate whether hedging aligns with the inaccurate component of a caption, similar in spirit to SPICE-style scene-graph evaluation (Anderson et al., 2016), could steer models toward hedging that is relevant and precise rather than "globally" present. Jung et al. establish that BLV users need information beyond object identification — context, brand identity, safety-relevant detail (Jung et al., 2022). Extending this to hedge placement, fine-tuning should target hedges on the attributes that carry the highest practical stakes for the user's task, rather than treating all caption elements as equally weighted.

Many of the conditions most likely to produce errors are also the least likely to produce an uncertainty signal. Models do already produce image-quality-aware language in some outputs — phrases like "the label is partially obscured" or "the image is unclear" appear in

the corpus — but inconsistently and without reliable association to caption error. Models should produce explicit statements about image quality when degradation is present, tied to specific conditions rather than left to emerge from training. Training examples can be constructed in which specific quality degradations are paired with explicit quality statements and appropriately scoped hedges. "This image appears blurry; I may not be reading the label correctly" is a disciplined version of something some models already do. The implication here is beyond making hedging more calibrated (a model-side intervention), but extends to making the model's limitations more transparent. If certain image quality predicts more caption error, models should *consistently* communicate those image quality problems explicitly as direct statements the user can act on. The work contributed by this thesis, which links image quality annotations to hedging behavior and caption accuracy on the same images, provides a foundation for fine-tuning toward this end.

Build precise hedging benchmarks.

Evaluating how models hedge, however, depends on having the means to measure whether they are doing it well — which current benchmarks do not provide. Disability-centered accounts of explainable AI argue that the criterion for a useful explanation is whether a user can act on it (Zhao et al., 2024). In this sense, the design of datasets used to evaluate and develop VLMs directly shapes whether model outputs support that kind of actionable understanding for BLV users. Prior research has called for more prevalent disability-centered model evaluation, arguing that evaluations built around BLV people's real conditions produce findings that standard benchmarks miss (Garg et al., 2026; Sharma et al., 2023). This thesis extends this call, as not only are models less accurate on the images BLV users actually take, they are less reliable in the communication they output. Communicative reliability, distinct from accuracy, has not been part of standard VLM evaluation. Adding it requires asking not just "does the model get this right?" but "does the model signal when it might not?" and "is that signal placed where it would help?'.

To this end, evaluating communication reliability requires a shift from caption-level to claim-level evaluation, where each caption is decomposed into structured components (e.g., product, brand, variety, and attributes) and each component is assessed independently. For every claim, evaluation should capture not only whether it is correct, but whether it should be hedged given the available visual evidence, and whether the model in fact hedges it. This enables direct measurement of hedge placement, distinguishing between correct-but-hedged (overcautious), wrong-and-unhedged (misleading), and wrong-and-hedged (aligned) outputs.

Because image quality is a known driver of model error, evaluation should test whether uncertainty expression is responsive to these conditions, increasing under degradations such as blur or obstruction.

Collectively, these design criteria move towards communicative reliability. This grounds the evaluation of whether models not only produce accurate captions, but signal their limitations in ways that affords BLV users to be able to appropriately decide to trust, verify, or disregard a caption based on its language.

5.3 Limitations and Directions for Future Work

Several constraints limit these findings. First, the analysis is cross-sectional and correlational, meaning that the co-occurrence of hedging and inaccuracy, and the association between blur and hedging, cannot be taken as evidence of causal effect. Both hedging and inaccuracy are properties of the same model output, and both may trace to the model's processing of a difficult image, instead of one driving the other. Disentangling these relationships would require experimental designs in which image quality, prompt structure, and model state are manipulated independently.

Furthermore, VLM outputs are stochastic so the captions analyzed here represent a single sample per image per model under fixed generation parameters. Hedge presence or absence

for a given image may not be stable across repeated generations. Future work incorporating multiple samples per image would allow analysis of hedging reliability, and whether stability of hedging across generations is itself informative about model confidence.

Next, the hedge detection pipeline relied on rule-based heuristics covering a defined set of hedging language. This approach is scalable and reproducible but is likely to miss contextual, multi-word, or embedded expressions of uncertainty, and may miss uncommon hedges not within the heuristic.

The most consequential limitation is the absence of BLV user data. All inferences about how hedging affects user trust, verification behavior, and decision-making are grounded in prior work (Huh et al., 2024; Karamolegkou et al., 2024) and linguistic theory (Lakoff, 1973), not in direct evidence from this study's population of interest. Whether the patterns documented here translate into meaningfully different user experiences — whether a hedged caption actually leads a BLV user to seek confirmation, and whether a non-hedged incorrect caption actually misleads — remains an empirical question. A user study in which BLV participants interact with captioning systems under varying image quality conditions, with and without hedged outputs, is the necessary and most important next step.

5.4 Conclusion

For a BLV user, navigating uncertainty is a normative process in visual access yet current model evaluation does not consider their verification preferences. This thesis has measured one dimension of how current models fail to support users in that condition: hedging language that does not track model reliability, that appears at rates driven by training data rather than visual uncertainty, and that is absent or inconsistent in the conditions that most demand it.

How models communicate the limits of what they know is a design choice, and the current default, communicative behavior as a training artifact, is not an effective one for the users

who depend on it. Drawing on the concept of everyday uncertainty (Tang et al., 2025), this thesis calls for evaluation frameworks and training pipelines that center communicative reliability seriously alongside accuracy. This means measuring hedge accuracy, not just hedge presence, while also including hedging as an explicit training objective for accessibility applications. It means designing captions for the user who cannot independently verify what a model is outputting, which is to say, designing them as if the stakes of getting the communication right are as high as the stakes of getting the fact right.

The dataset and annotation approach contributed here offer a perspective for measuring that communication, and the design questions raised, about calibrated hedging, explicit image quality reporting, model-specific user guidance, and the user studies needed to evaluate all of the above, point towards next steps for building assistive captioning systems that support informed decision making and strategies for communicating uncertainty.

Bibliography

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Rudaiba Adnin and Maitraye Das. "i look at it as the king of knowledge": How blind people use and understand generative ai tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706776. doi: 10.1145/3663548.3675631. URL https://doi.org/10.1145/3663548.3675631.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N Brewer. Misfitting with ai: How blind people verify and contest ai errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–17, 2024.

AllenAI. Allenai/Molmo-72B-0924 · Hugging Face. https://huggingface.co/allenai/Molmo-72B-0924, 2024.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

Be My Eyes. Introducing: Be My AI, August 2023.

Andy Beatman and Ailsa Leen. 6 ways generative AI helps improve accessibility for all with Azure, March 2024. URL https://azure.microsoft.com/en-us/blog/6-ways-generative-ai-helps-improve-accessibility-for-all-with-azure/.

David M. Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A. Ross, and John Canny. IC3: Image Captioning by Committee Consensus. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.556.

Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, pages 1–18, New York, NY, USA, October 2024. Association for Computing Machinery. ISBN 979-8-4007-0628-8. doi: 10.1145/3654777.3676375.

Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing Image Quality Issues for Real-World Problems. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653, June 2020. doi: 10.1109/CVPR42600.2020.00370.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960. URL `https://api.semanticscholar.org/CorpusID:15926286`.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 91–104, 2025.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: K-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 7750–7774, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A Young, and Brian Belgodere. Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge. *Journal of Artificial Intelligence Research*, 73:437–459, 2022.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Kapil Garg, Xinru Tang, Jimin Heo, Dwayne R Morgan, Darren Gergle, Erik B Sudderth, and Anne Marie Piper. "it's trained by non-disabled people": Evaluating how image quality affects product captioning with vision-language models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY,

USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791309. URL https://doi.org/10.1145/3772318.3791309.

Google. Gemini 2.5 Flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL http://arxiv.org/abs/2407.21783.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, 2024.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12362, pages 417–434. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-58520-4_25. URL https://doi.org/10.1007/978-3-030-58520-4_25.

Kenta Hama, Takashi Matsubara, Kuniaki Uehara, and Jianfei Cai. Exploring uncertainty measures for image-caption embedding-and-retrieval task. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(2):1–19, 2021.

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. Long-form answers to visual questions from blind and low vision people. In *First Conference on Language Modeling*, 2024.

Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S Ackerman. "so what? what's that to do with me?" expectations of people with visual impairments for image descriptions in their personal photo activities. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1893–1906, 2022.

Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5839–5849. Association for Computing Machinery, 2017. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025899. URL https://dl.acm.org/doi/10.1145/3025453.3025899.

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sánchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. Evaluating multimodal language models as visual assistants for visually impaired users. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658941. URL https://doi.org/10.1145/3630106.3658941.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

Cindy Kaiying Lin and Steven J. Jackson. From bias to repair: Error as a site of collaboration and negotiation in applied data science work. *Proc. ACM Hum.-Comput. Interact.*, 7 (CSCW1), April 2023. doi: 10.1145/3579607. URL https://doi.org/10.1145/3579607.

Meta. Meta-llama/Llama-3.2-90B-Vision-Instruct · Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct, December 2024.

Meta. Introducing the Meta AI App: A New Way to Access Your AI Assistant, April 2025.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Ananya Gubbi Mohanbabu and Amy Pavel. Context-Aware Image Descriptions for Web Accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, pages 1–17. Association for Computing Machinery, 2024. ISBN 979-8-4007-0677-6. doi: 10.1145/3663548.3675658. URL https://dl.acm.org/doi/10.1145/3663548.3675658.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 22047–22069, Red Hook, NY, USA, December 2023. Curran Associates Inc.

OpenAI. ChatGPT. https://chatgpt.com/, 2025a.

OpenAI. GPT-4.1. https://platform.openai.com/docs/models/gpt-4.1, 2025b.

OpenAI. Prompt engineering - OpenAI API, 2025c. URL https://platform.openai.com/docs/guides/prompt-engineering.

Anna Prokofieva and Julia Hirschberg. Hedging and speaker commitment. In *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*, 2014.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. Humans overrely on overconfident language models, across languages. *arXiv preprint arXiv:2507.06306*, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Ligao Ruan, Giles Hamilton-Fletcher, Mahya Beheshti, Todd E Hudson, Maurizio Porfiri, and John-Ross Rizzo. A multimodal assistive system for product localization and retrieval for people who are blind or have low vision. *arXiv preprint arXiv:2601.12486*, 2026.

Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. Disability-first design and creation of a dataset showing private visual information collected with people who are blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

Abigale Stangl, Kristina Shiroma, Nathan Davis, Bo Xie, Kenneth R. Fleischmann, Leah Findlater, and Danna Gurari. Privacy concerns for visual assistance technologies. *ACM Trans. Access. Comput.*, 15(2), May 2022. ISSN 1936-7228. doi: 10.1145/3517384. URL https://doi.org/10.1145/3517384.

Abigale Stangl, Emma Sadjo, Pardis Emami-Naeini, Yang Wang, Danna Gurari, and Leah Findlater. "Dump it, Destroy it, Send it to Data Heaven": Blind People's Expectations for Visual Privacy in Visual Assistance Technologies. In *Proceedings of the 20th International Web for All Conference*, W4A '23, pages 134–147. Association for Computing Machinery, 2023. ISBN 979-8-4007-0748-3. doi: 10.1145/3587281.3587296. URL https://dl.acm.org/doi/10.1145/3587281.3587296.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Xinru Tang, Ali Abdolrahmani, Darren Gergle, and Anne Marie Piper. Everyday uncertainty: How blind people use genai tools for information access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713433. URL https://doi.org/10.1145/3706598.3713433.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem

Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord, editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

Dell Zhang, Murat Sensoy, Masoud Makrehchi, Bilyana Taneva-Popova, Lin Gui, and Yulan He. Uncertainty quantification for text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3426–3429, 2023.

Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. VIALM: A survey and benchmark of visually impaired assistance with large models. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA, December 2023. Curran Associates Inc.

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

A Appendix

A.1 Dataset

Listing A.1: Hedge detection regular expressions.

```
hedge_patterns = [
    # Epistemic verbs
    r'\b(appears?␣to␣be|seems?␣to␣be|looks?␣like|seems?␣like)\b',
    r'\b(possibly|probably|perhaps|maybe|might|could␣be|may␣be)\b',
    r'\b(likely|unlikely|presumably|apparently|seemingly)\b',
    r'\b(suggests?|implies?)\b',
    # Approximators and rounders
    r'\b(approximately|roughly|about|nearly|almost)\b',
    r'\b(some␣kind␣of|some␣sort␣of|some␣type␣of)\b',
    r'\b(or␣so|more␣or␣less)\b',
    # Vagueness
    r'\b(something|somewhere|somewhat)\b',
]
```

Appendix: Hedging Code-book for Product Captions

Coding: All captions presented contain language that *could* be uncertain. Identify whether the hedges used within a caption indicates approximation or speculation about the main product or object of interest. Each caption should be coded as **HEDGE (1)** or **NOT A HEDGE (0)**.

Annotation Rules

- **HEDGE (1)**: Captions signaling uncertainty about the product using words such as "likely", "maybe", "appears", "probably", "sort of", "approximately" , etc.
- **NOT A HEDGE (0)**: Captions with definite statements about the product or hedging limited to background/non-critical context.
- Compound sentences: If any part hedges about the product, code the entire sentence as HEDGE.

Caption	Code
Plastic Gatorade bottle, clear contents, suggesting **likely** lemon-lime or other clear flavor.	1
Can's color is blue, shape cylindrical; variety information unclear.	0
Product name in silver text, variety in green and white; background **appears to be** a bathroom.	0
Red Jimmy Dean box; box shows three sausage patties with what **appears to be** an omelet. Box states "FULLY COOKED."	1
Can of cream of mushroom soup, **likely** store or generic brand, with red and white label.	1
Box of Select Choice granola bars, 5 bars; box **mostly** red with picture of bar containing oats and chocolate chips.	1
Plastic spray bottle with orange trigger; background **appears to be** a carpeted room.	0
Glass jar of Heinz pickles, partial red text visible; glass **appears** empty or **almost** empty.	1
Ocean Spray White Cran Peach juice bottle; clear, empty or **nearly** empty. Blue oval logo "Ocean Spray". Orange banner "White Cran·Peach".	1

Table A.1: Hedging examples and coding from dataset.

Example Captions and Coding

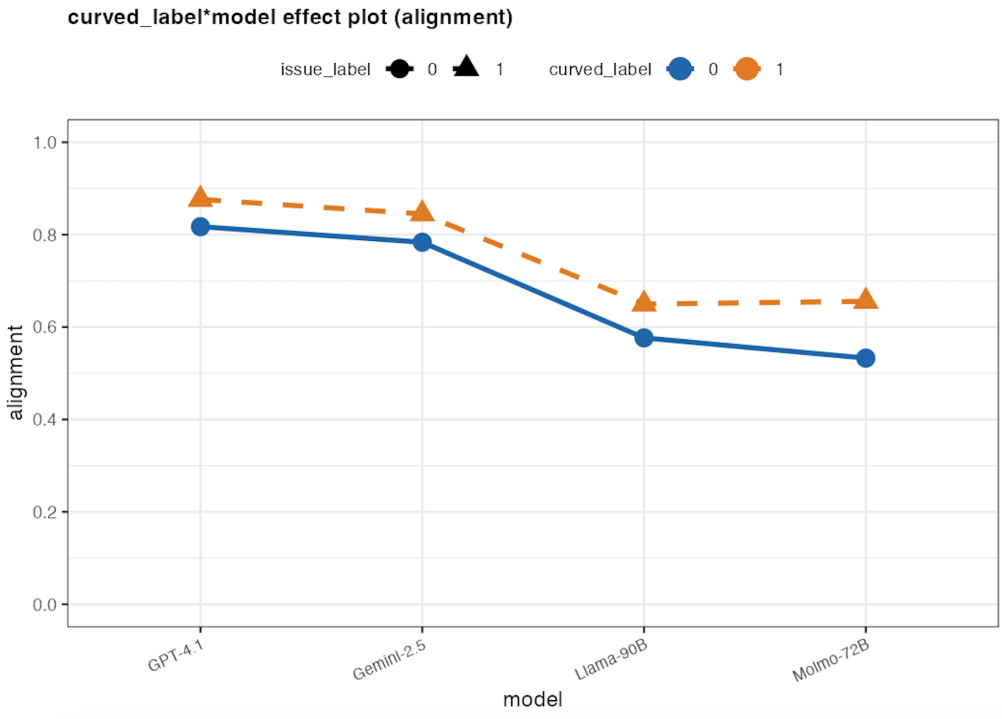

Figure A.1: Curved Label alignment Per Model.

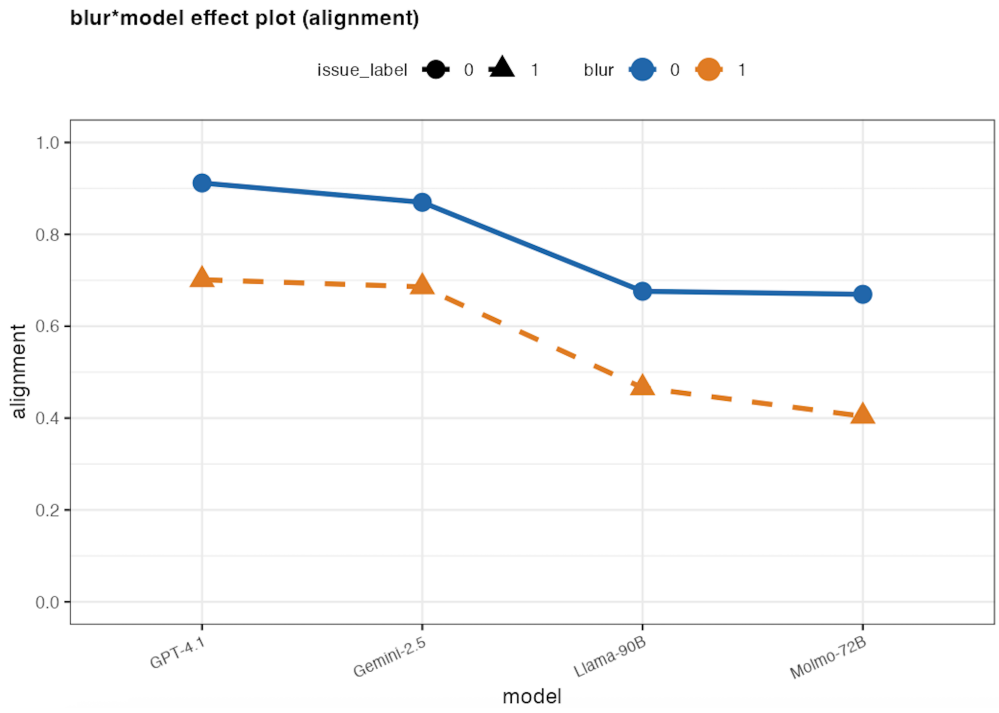

Figure A.2: Blur alignment Per Model.

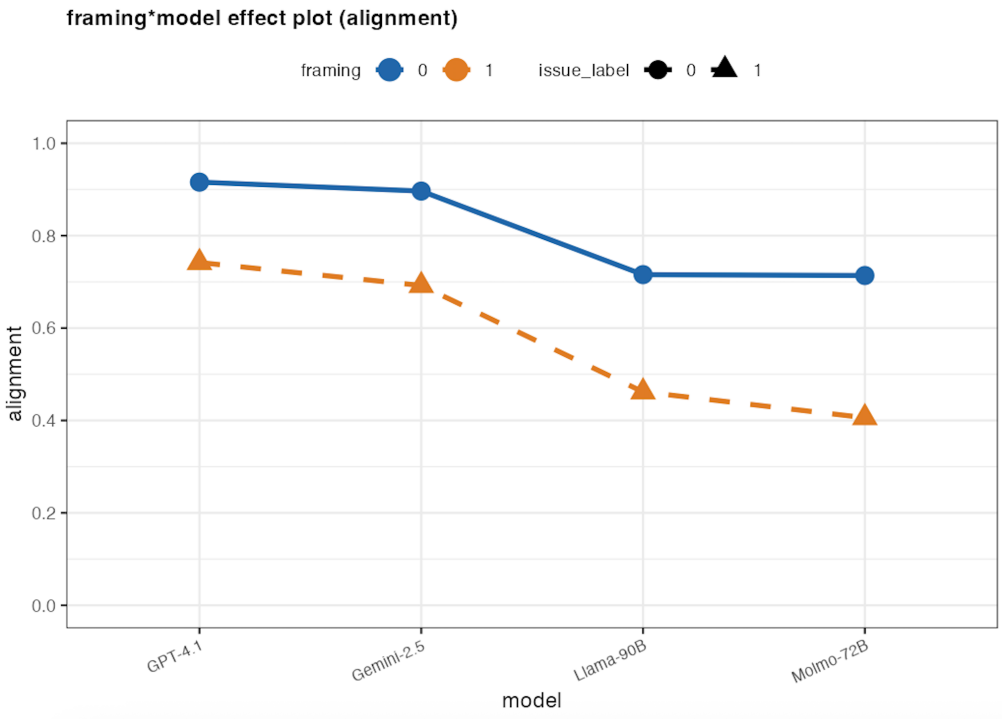

Figure A.3: Framing alignment Per Model.

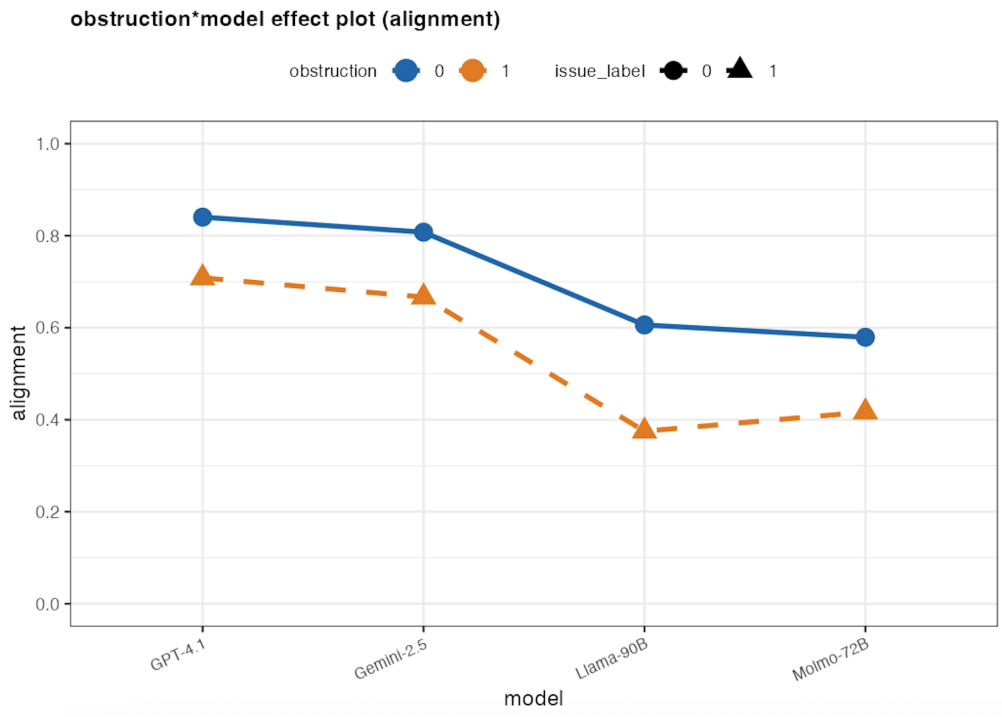

Figure A.4: Obstruction alignment Per Model.

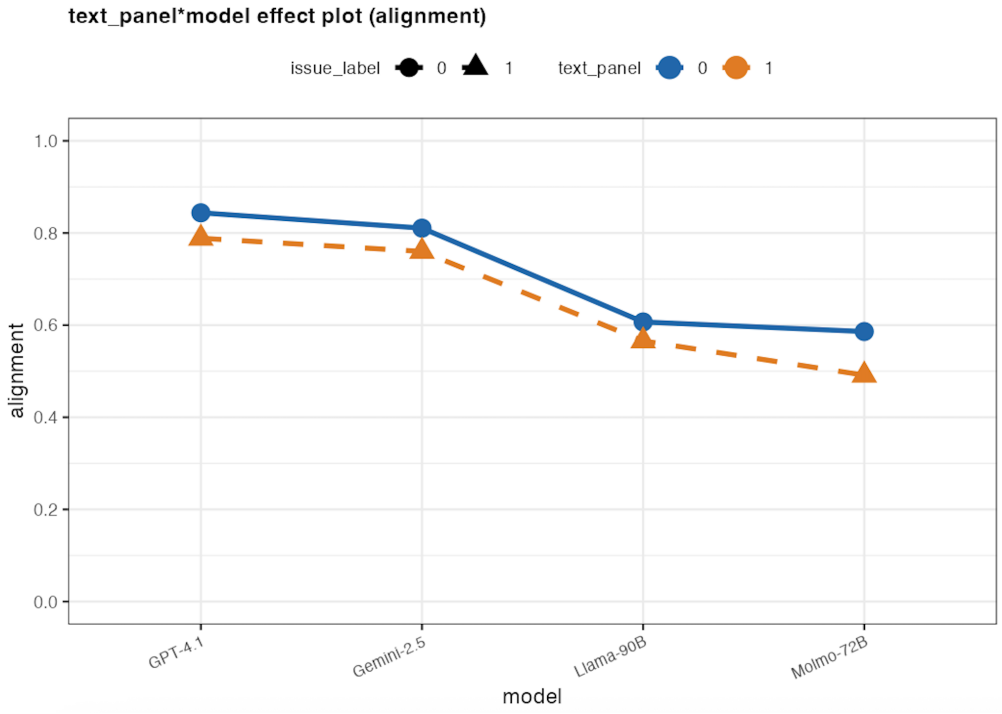

Figure A.5: Text Panel alignment Per Model.

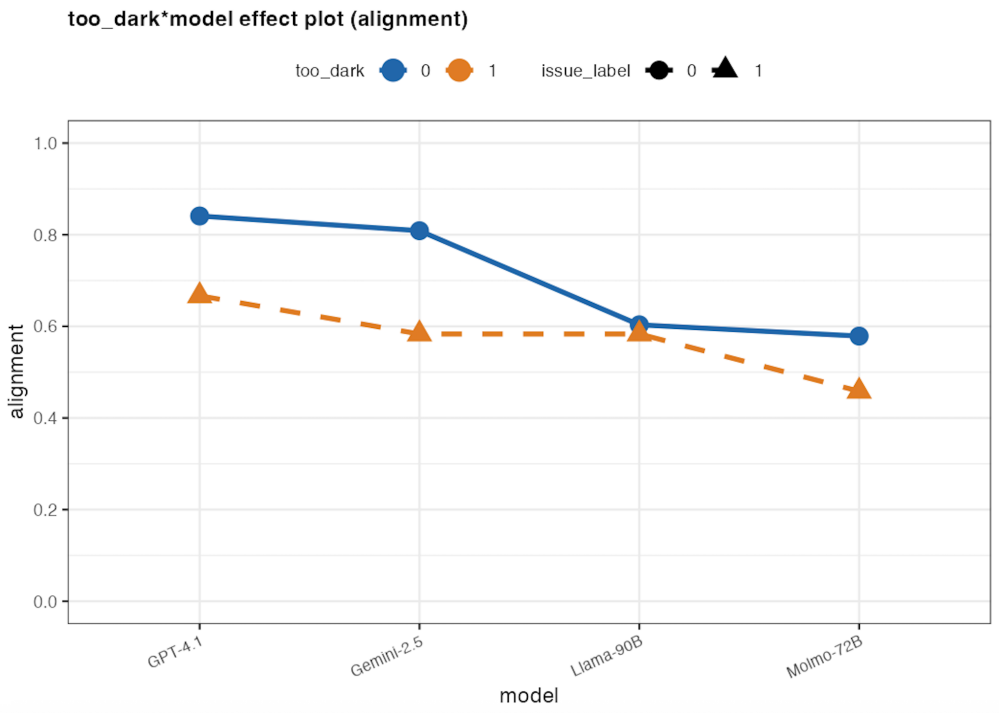

Figure A.6: Blur alignment Per Model.

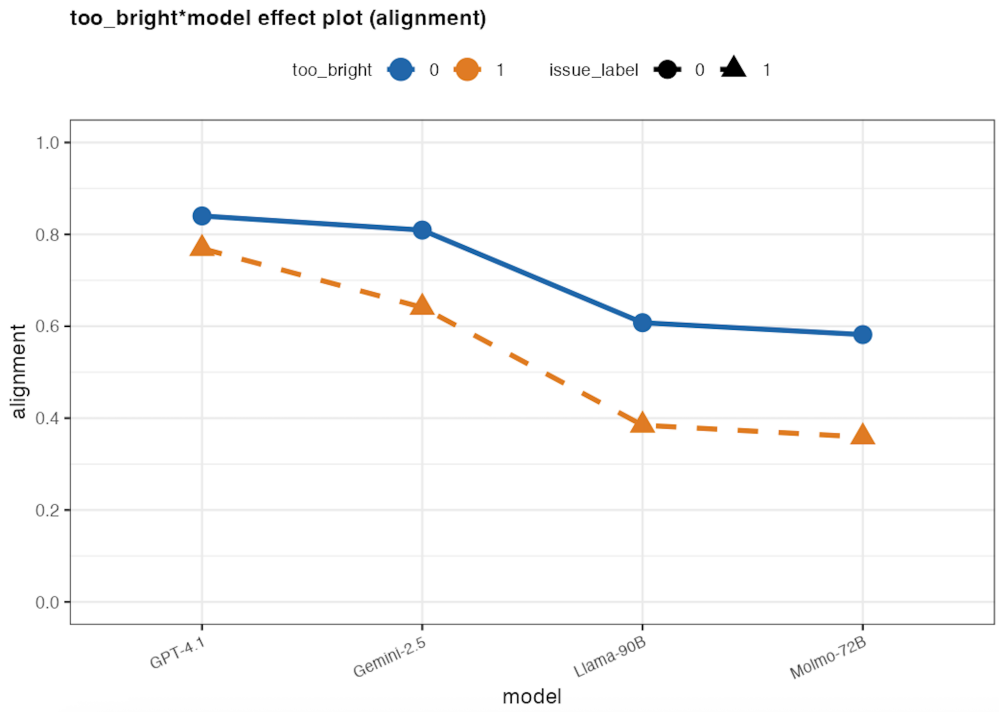

Figure A.7: Blur alignment Per Model.

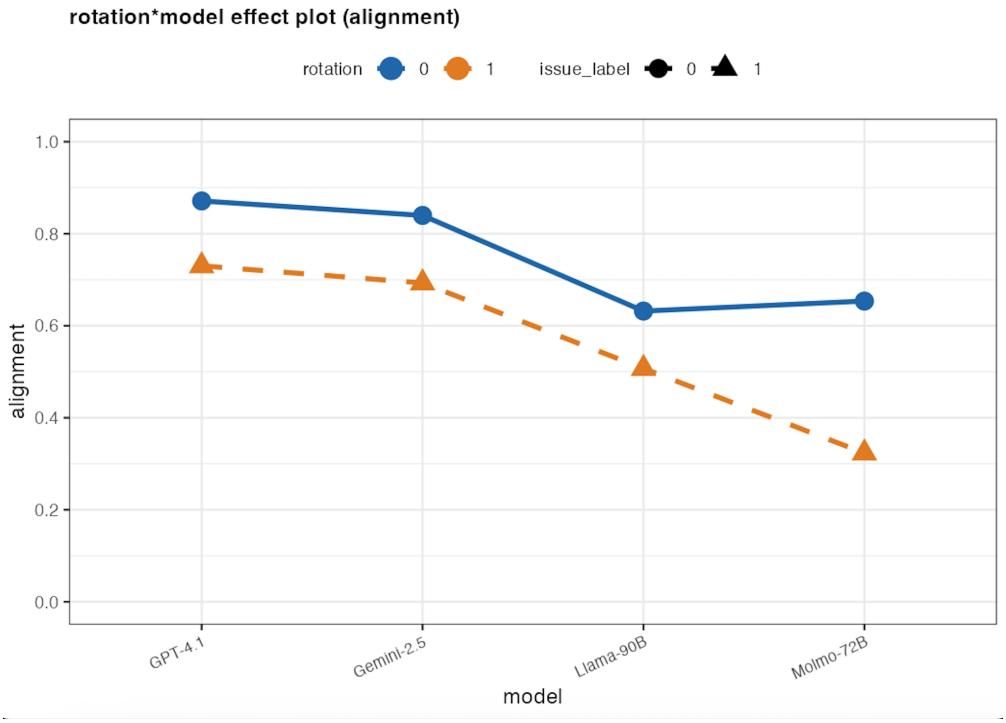

Figure A.8: Obstruction alignment Per Model.

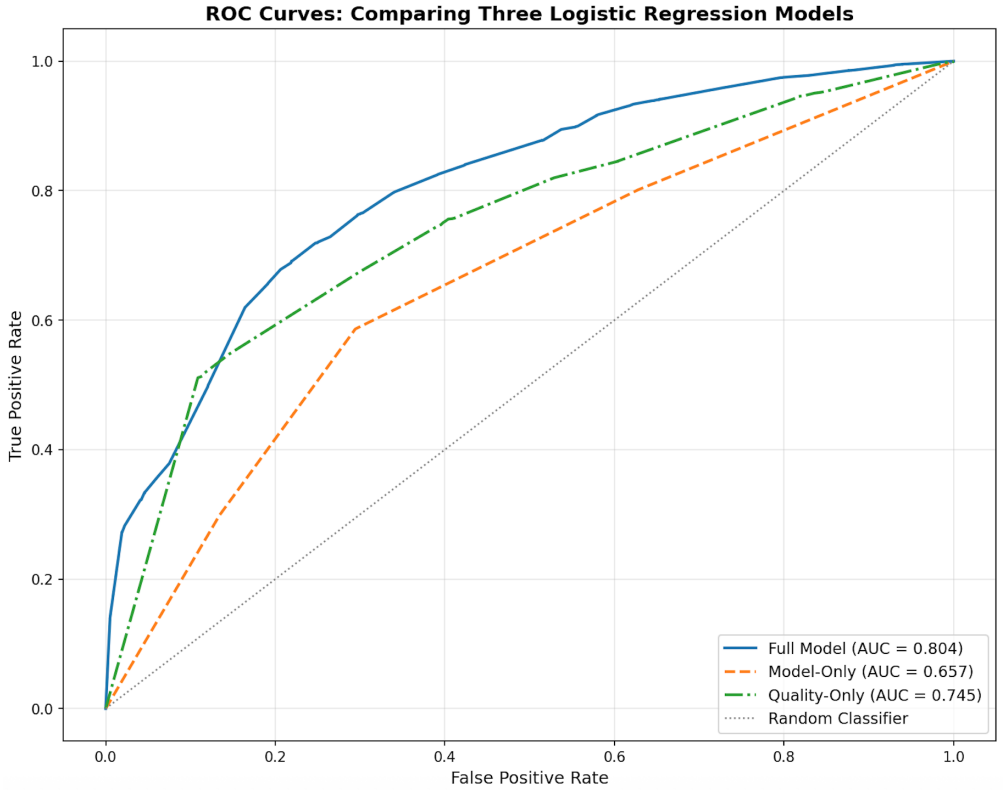

Figure A.9: Obstruction alignment Per Model.