PARC: A Quantitative Framework Uncovering the Symmetries within Vision Language Models

Supplementary Material

VLM Name	Huggingface ID	Model Size	LLM Size	Input Resolution	Training Data	Data Cur.
LLaVA-1.5 7B	liuhaotian/llava-v1.5-7b	7.4B	7.0B	336×336	1.2M	high
LLaVA-1.5 13B	liuhaotian/llava-v1.5-13b	13.4B	13.0B	336×336	1.2M	high
LLaVA-1.6 7B vic	liuhaotian/llava-v1.6-vicuna-7b	7.1B	6.7B	336×336	1.6M	high
LLaVA-1.6 7B mis	liuhaotian/llava-v1.6-mistral-7b	7.3B	7.0B	336×336	1.6M	high
LLaVA-1.6 13B vic	liuhaotian/llava-v1.6-vicuna-13b	13.3B	13.0B	336×336	1.6M	high
LLaVA-1.6 34B	liuhaotian/llava-v1.6-34b	34.7B	34.4B	336×336	1.6M	high
Qwen-VL	Qwen/Qwen-VL	9.6B	7.7B	224×224	1400.0M	low
Qwen-VL Chat	Qwen/Qwen-VL-Chat	9.6B	7.7B	224×224	1400.0M	low
CogVLM Chat	THUDM/cogvlm-chat-hf	*d.o.*	7.0B	*d.o.*	2250.0M	low
CogAgent	THUDM/cogagent-vqa-hf	*d.o.*	7.0B	*d.o.*	1690.0M	low
CogVLM GG	THUDM/cogvlm-grounding-generalist-hf	*d.o.*	7.0B	*d.o.*	1500.0M	low
CogVLM2 19B	THUDM/cogvlm2-llama3-chat-19B	19.0B	8.0B	1344×1344	*n.a.*	*n.a.*
InternVL2 1B	OpenGVLab/InternVL2-1B	0.8B	0.5B	448×448	>50M	*n.a.*
InternVL2 2B	OpenGVLab/InternVL2-2B	2.5B	2.1B	448×448	>50M	*n.a.*
InternVL2 4B	OpenGVLab/InternVL2-4B	4.1B	3.8B	448×448	>50M	*n.a.*
InternVL2 8B	OpenGVLab/InternVL2-8B	8.1B	7.7B	448×448	>50M	*n.a.*
InternVL2 26B	OpenGVLab/InternVL2-26B	25.5B	19.9B	448×448	>50M	*n.a.*
InternVL2 40B	OpenGVLab/InternVL2-40B	40.1B	34.4B	448×448	>50M	*n.a.*
Cambrian 3B	nyu-visionx/cambrian-phi3-3b	*d.o.*	3.3B	384×384	9.5M	high
Cambrian 8B	nyu-visionx/cambrian-8b	*d.o.*	8.0B	384×384	9.5M	high
Cambrian 13B	nyu-visionx/cambrian-13b	*d.o.*	13.0B	384×384	9.5M	high
Cambrian 34B	nyu-visionx/cambrian-34b	*d.o.*	34.4B	384×384	9.5M	high

Table A1. Summary of analyzed VLMs in PARC and their properties. "Data Cur." means Data Curation. The huggingface ID was used in our evaluation code to load the respective models. VLM properties are listed according to the corresponding papers or their github repositories. Note that some information marked with "*n.a.*" is not publicly available, while "*d.o.*" indicates that the information is difficult to obtain. The training data estimate for InternVL2 is based on its predecessor InternVL 1.2 plus with 39.3M pretraining and 12M finetuning data. The data curation estimate is "high" if the training data is a balanced collection of curated or established datasets for specific tasks, and "low" for web-crawled, auto-annotated data.

A1. Evaluated Vision Language Models

In Tab. A1 we give an overview of all 22 evaluated VLMs and their model properties, which form the basis for our experimental evaluation in Sec. 3. The 22 models come from the 7 VLM families LLaVA1.5, LLaVA1.6, Qwen-VL, CogVLM, CogVLM2, InternVL2 and Cambrian [3, 4, 6, 11, 12, 17, 18].

A2. PARC Prompt Variations and Datasets

Below we give more details on the generation of our prompt variations and datasets. First, we describe our prompt variation taxonomy in greater detail. Next, we provide the LLaMA3 prompts [1], which were used to generate the language variations across all datasets. Then, we give details on creating the semantic vision exchange variation (VS-E), which requires manual data annotation and results in 6 new datasets. Finally, we describe how we cleaned the prompts in the MMBench dataset [13] to enable our language prompt alterations with LLaMA3.

A2.1. Prompt Variation Taxonomy

We define any change to a prompt's text or vision component as prompt variation, focusing on human-plausible variations that reformulate (same expected answer) or semantically change (changed expected answer) the prompt. PARC proposes expanding *semantic changes* to images, the remaining vision and language *reformulations* in PARC are a subset of realistic, existing text and image *perturbations*:

$$\text{reformulation}\ {}^{\text{vision}}_{\text{language}} \subset \text{perturbation}\ {}^{\text{image}}_{\text{text}} \subset \text{prompt variation}.$$

We choose them to cover broad perturbation classes, established in prior works. For language reformulations, we adopted several text perturbation classes from [15]: AddText [LR-V Verbose], DropText [LR-C Concise], and TextStyle [LR-I Instruction]; other classes in [15] that yield syntactically incorrect texts are omitted in PARC because they are unlikely to be made by real-world users. For vision reformulations, we adopt major image perturbation classes from [5]: Color [VR-L Lighting], Detail change [VR-B Blur], and Spatial changes [VR-R Rotation]. We selected three representative reformulations for language/vision as we also consider computational feasibility for VLM evaluations: 11 total variations on 7 datasets already lead to extensive 84 dataset evaluations *per VLM*. Therefore, three representatives of diverse perturbation classes maximize prompt variability under evaluation budget constraints.

A2.2. Language Variations: LLaMA3 Prompts

For our language prompt variations, we use LLaMA3-70B [1] to alter the language component of the original prompts. Below, we provide the complete prompts that were used with LLaMA3 to alter an original text prompt, the `<Prompt to alter>`, into the variants Instruction (LR-I), Concise (LR-C), Verbose (LR-V), Not (LS-N), Antonyms (LS-A) and More-Less (LS-M).

LR-I – Instructions.

```
Rewrite the given phrases into instructions. Keep the instruction short, and as
close to the original sentence as possible. If the phrase is already an
instruction, keep the original phrase. Only return a single new instruction,
do NOT give additional explanations.

Q: "Are the two sofas the same color in the picture?"
A: "Determine if the two sofas are the same color in the picture."

Q: "Based on the image, how can fun and engaging toothbrush holders help
children develop better dental health habits?"
A: "Based on the image, identify how fun and engaging toothbrush holders can
help children develop better dental health habits."

Q: "Which pants' fit is more regular?"
A: "Determine which pants' fit is more regular."

Q: "In the picture, which direction is the teddy bear facing?"
A: "Identify the direction in which the teddy bear is facing in the picture."

Q: "<Prompt to alter>"
A:
```

LR-C – Concise.

```
Rephrase the following question using fewer words. Make sure the meaning of the
question stays the same. Return the same question if it is not possible to use
fewer words. For your context, the question is about an image that is not provided
here. Ensure correct grammar. Only return a single new question, do NOT give
additional explanations. Use fewer words than the original question.

Q: "Which road is more paved?"
A: "More paved road?"

Q: "Are the two sofas the same color in the picture?"
A: "Are the sofas the same color?"

Q: "Which cloud is whiter?"
A: "Whiter cloud?"

Q: "Which jacket is more asymmetrical??"
A: "More asymmetrical jacket?"

Q: "<Prompt to alter>"
A:
```

LR-V – Verbose.

```
Rephrase the following question to make it more verbose. Make sure the meaning of
the question stays the same, do NOT invent additional details. For your context,
the question is about an image that is not provided here. Ensure correct grammar.
Only return a single new question, do NOT give additional explanations.

Q: "<Prompt to alter>"
A:
```

LS-N – Not.

```
Change the following questions to ask for the opposite by negating them with 'not'.
Keep the question as close to the original as possible. Ensure correct grammar.
Return a single, complete question. Do NOT abbreviate parts of the question.
Do NOT give additional explanations.

Q: "Which road is more paved?"
A: "Which road is not more paved?"

Q: "Which umbrella is yellower?"
A: "Which umbrella is not yellower?"

Q: "Which cat is staring more directly at the camera?"
A: "Which cat is notstaring more directly at the camera?"

Q: "Which fence is burnter?"
A: "Which fence is not more burnt?"

Q: "<Prompt to alter>"
A:
```

LS-A – Antonyms.

```
Change the following questions to ask for the opposite by using antonyms or words
with opposite meanings. Keep the question as close to the original as possible.
Ensure correct grammar. Return a single, complete question. Do NOT abbreviate parts
of the question. Do NOT give additional explanations.

Q: "Which road is more paved?"
A: "Which road is more unpaved?"

Q: "Which cat is staring more directly at the camera?"
A: "Which cat is looking more away from the camera?"

Q: "Which bike is more folded?"
A: "Which bike is more unfolded?"

Q: "In what direction is Chile from Peru?"
A: "In what direction is Peru from Chile?"

Q: "Which Python code can generate the content of the image?"
A: "Which Python code refuses generating the content of the image?"

Q: "What musn't Joe and Alice trade to each get what they want?"
A: "What musn't Joe and Alice trade to each get what they want?"

Q: "Which apple is closer to the camera?"
A: "Which apple is farther from the camera?"

Q: "Which bottle is more dented?"
A: "Which bottle is more intact?"

Q: "Which aluminum is more molten?"
A: "Which aluminum is more solid?"

Q: "<Prompt to alter>"
A:
```

LS-M – More-Less.

```
Change the following questions to ask for the opposite by switching the word 'more'
for 'less' or 'fewer'. Keep the question as close to the original as possible.
Ensure correct grammar. Return a single, complete question. Do NOT abbreviate parts
of the question. Do NOT give additional explanations.

Q: "Which road is more paved?"
A: "Which road is less paved?"

Q: "Which umbrella is yellower?"
A: "Which umbrella is less yellow?"

Q: "Which bed is closer to the camera?"
A: "Which bed is less close to the camera?"

Q: "Which animal has longer hair?"
A: "Which animal has less long hair?"

Q: "Which cat is staring more directly at the camera?"
A: "Which cat is staring less directly at the camera?"

Q: "Which fence is burnter?"
A: "Which fence is less burnt?"

Q: "<Prompt to alter>"
A:
```

A2.3. Vision Variations

Not all datasets supports every semantic vision variations that are proposed in Tab. 1. Below, we describe how we create the visual semantic exchange variation (VS-E), and give an overview of the resulting datasets after this manual annotation process.

VS-E – Image Exchange. The visual semantic exchange variation is a semantic change, and therefore needs to modify the correct answer by changing the image component of the prompt. The idea is as follows: For a prompt with two images (left and right) and a question of the type "Which image is more ...?", we exchange the image that has more of this attribute, *i.e*. that *wins* the comparison, for one that has much less of this attribute, *i.e*. that *loses* the comparison. For example, in Tab. 1, we ask "Which animal has longer fur?". Here, the alpaca on the left wins, because it has longer fur than the otter on the right. Now we exchange the alpaca on the left with an animal that has less long fur than the otter on the right, such that the otter on the right becomes the winner.

In practice, we identify valid replacement images in the following way *per dataset* $D = \{p_n\}_{n=1}^{N}$ that is composed of prompts $p = (I^w, I^l, T)$, which are triplets of a winning image I^w, a losing image I^l and a text T:

1. Identify unique question texts within the dataset, and collect all instances of prompts with those unique question texts. *E.g*., collect all instances of prompts that ask "Which cloud is more white?". Only keep prompts belonging to a unique question with at least two instances.
2. From the prompts for one unique question text, collect all *losing* images I^l, *e.g*. all "less white clouds".
3. Then, per losing image, mark all losing images this image would win against as possible exchanges. *E.g*. mark dark gray clouds as possible exchanges when considering a lighter gray cloud.
4. Go through the prompts for the unique question text again. Per prompt, check if the losing image has an annotated image it would win against. If so, exchange

	M-S	M-A	V-S	V-A	NYU	Fas
Number of Data Samples						
Initial Samples	197	597	940	470	1858	2435
VS-E Filtering	56	112	108	71	385	114
VS-E Percentage	28.4%	18.8%	11.5%	15.1%	20.7%	4.7%
LLaVA-1.6 34B Error Rates						
Initial	89.7	84.9	79.3	77.7	67.2	72.1
Filtered	87.5	85.7	86.1	92.5	70.4	71.9

Table A2. Dataset statistics for the datasets MIT-States (M-S), MIT-Attributes (M-A), VAW-States (V-S), VAW-Attributes (V-A), NYU-Depth V2 (NYU) and Fashionpedia (Fas) [7, 8, 14, 16] as used in PARC. The *Initial* comparative prompts are as curated in CompBench [9], which uses 20% of the *Initial* data for evaluations. Our additional data annotation to enable Vision Semantic Changes by Exchanging images (VS-E) also retains about 20% of this data, on which we evaluate the prompt sensitivity with PARC. LLaVA-1.6 34B error rates on initial and filtered prompts.

the winning image with the annotated image. Discard prompts where the losing image does not have an annotated exchange.

This process also works if the role of winning and losing images is swapped. In the example from Tab. 1, we would then exchange the losing image (the otter) with one that wins against the alpaca, *e.g.* a mammoth. We identify unique question texts in the datasets MIT-States (M-S), MIT-Attributes (M-A), VAW-States (V-S), VAW-Attributes (V-A), NYU-Depth V2 (NYU) and Fashionpedia (Fas) [7, 8, 14, 16], and manually label potential exchange images per dataset.

Dataset Statistics after VS-E. The image exchange process described above reduces the number of available samples, because prompts are discarded if their text does not have multiple occurrences or if no suitable exchange image is found. In Tab. A2 we show the number of initial prompts, and how many prompts are retained after discarding those whose images cannot be exchanged. We retain about 20% of the prompts, which is similar to the evaluation of those datasets in [9], which randomly splits these datasets into 20% test and 80% training data.

We also report accuracies for LLaVA-1.6 34B on the original and filtered datasets. They are generally well aligned, except for VAW-derived data. For V-S and V-A, the filtered datasets are slightly easier, as filtering for exchangeable questions VS-E discards rare, difficult questions with single occurrences, which are more prominent on VAW. We refrained from further filtering on these two datasets to maintain adequate sample size for subsequent evaluations. Note that this does not affect insights from PARC, as all models are evaluated on the same data.

A2.4. MMBench Data Cleaning Protocol

We use the prompts of the MMBench [13] val split to create multiple prompt variations within PARC. Because MMBench is composed from multiple data sources, the prompts do not follow a uniform structure. To enable a consistent modification with our prompt variation framwork and ensure that the original prompts are sensical, we manually clean the 382 unique text prompts using the following steps:

1. Remove unrelated sentence parts with convoluting information that is irrelevant to the question and its answer.
2. Unify multiple sentences into one.
3. Rephrase prompts to grammatically correct questions.
4. Unify sentence capitalization.
5. Where possible, remove any answer-leakages from questions.
6. Remove questions that do not support negations via antonym replacement.

For steps 2 and 3, we use LLaMA3-70B [1] to generate an initial set of proposed questions which are manually checked and corrected where necessary. The following prompt is used for the initial question rephrasing:

```
Rewrite the given phrases into questions. Keep the question as close to the
original sentence as possible. If the phrase is already a question, keep the
original question. Ensure correct grammar. Only return a single new question,
do NOT give additional explanations.

Q: "where is the bike?"
A: "Where is the bike?"

Q: "What direction is India in Kyrgyzstan?"
A: "In what direction is India from Kyrgyzstan?"

Q: "Based on the image, how can fun and engaging toothbrush holders help
children develop better dental health habits?"
A: "Based on the image, how can fun and engaging toothbrush holders help
children develop better dental health habits?"

Q: "Complete the sentence.
The African elephant is the () land animal in the world."
A: "How do you complete the following sentence? 'The African elephant is
the () land animal in the world.'"

Q: "Two magnets are places as shown. Will these magnets attract or repel
each other?"
A: "If the two magnets are placed as shown, will they attract or repel
each other?"

Q: "<Prompt to Rephrase>"
A:
```

In step 5, an example of removing answer-leakages from the following original prompt:

```
Based on the image, what does the dog's
behavior of jumping and playing Frisbee
indicate about its well-being?
(A) The dog is participating in a
    professional Frisbee competition.
(B) The dog is engaged in physical activity,
    promoting its health and well-being.
(C) The dog is attempting to catch
    a bird in mid-air.
(D) The dog is bored and looking for
    something to do.
```

This text question invalidates answers `C` and `D` without even seeing the image, because it mentions the dog interacts with a frisbee. After removal, the question becomes `Based on the image, what does the dog do?`

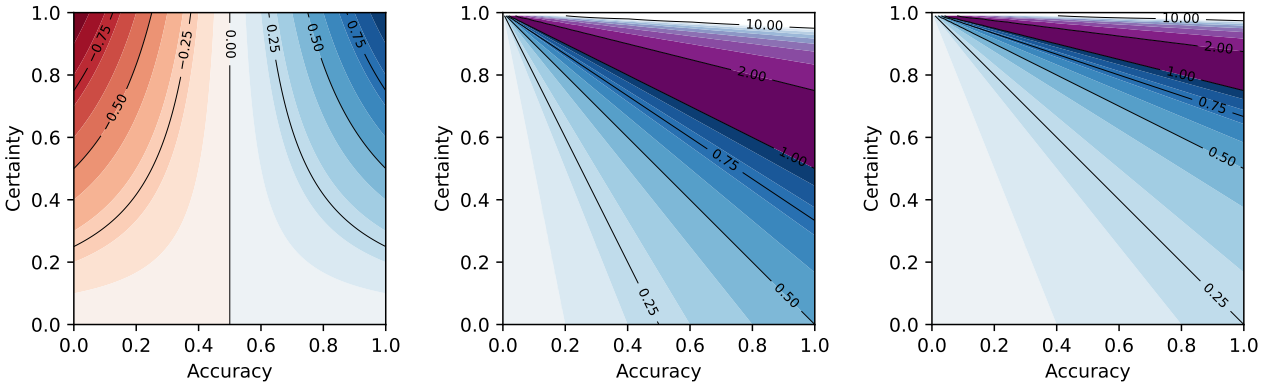

Figure A1. Mapping from certainty and accuracy to reliability score with $\text{acc}_{\text{rand}} = 0.5$ *[Left]* vs. uncertainty-aware accuracy UAcc with $|\mathcal{P}| = 4$ *[Middle]* and $|\mathcal{P}| = 16$ *[Right]*. Our reliability score provides two guarantees (one for accuracy and one for certainty) that can be directly seen from any score except 0. Uacc provides at most one guarantee for either accuracy or certainty. Further, its maximum changes with the number of answers per prompt $|\mathcal{P}|$, generating scores that are incomparable across datasets.

A3. Analysis of PARC's Reliability Score

Calibration. For calibration, the reliability should be 0 for $acc = acc_{rand}$, independent of $cert$. This condition implicitly calibrates the vanilla definition $rel = (2 \cdot acc - 1) \cdot cert$ to $acc_{rand} = 0.5$. Thus to calibrate, we generalize the reliability definition to $rel = (2 \cdot acc^m - 1) \cdot cert$ and solve $0 = (2 \cdot acc_{rand}^m - 1)$, obtaining $m = \frac{\log(1/2)}{\log(acc_{rand})} = \frac{2}{\log(1/acc_{rand})}$. For $acc_{rand} = 0.5$ this translates to $m = 1$ and yields the vanilla reliability, validating our generalization.

Comparison to Other Scores. The reliability score we propose with PARC is not the first to combine accuracy and uncertainty into a unified number. Hence, we compare the interpretability of our reliability against the uncertainty-aware accuracy UAacc from [10]. UAcc is introduced within a framework that uses conformal prediction [2, 19]. Therefore, its certainty notion also builds on the set sizes of the set of all possible answers $\mathcal{P}$ and the prediction set $\mathcal{C}$:

$$\text{UAcc} = \frac{\text{acc}}{|\mathcal{C}|}\sqrt{|\mathcal{P}|} \tag{1}$$

$$= \frac{\text{acc}}{(1 - \text{cert}) \cdot (|\mathcal{P}| - 1) + 1}\sqrt{|\mathcal{P}|}. \tag{2}$$

UAcc takes values in $[0, \sqrt{|\mathcal{P}|}]$, where the maximum exceeds 1 and depends on the number of possible answers $|\mathcal{P}|$. This makes it impossible to compare UAcc scores across datasets with varying answer numbers, because their score maximum is already inconsistent. Fig. A1 compares the mapping from certainty and accuracy to UAcc with $|\mathcal{P}| = 4$ and $|\mathcal{P}| = 16$ to our reliability score. And even within the same number of answers $|\mathcal{P}|$, UAcc does not offer truly meaningful built-in guarantees about the individual accuracies and certainties that can be seen from a single score at a glance: Any given score maps to the full spectrum of possible accuracies or certainties, *e.g.* for $|\mathcal{P}| = 4$, Uacc ≤ 0.5 may represent any certainty and Uacc ≥ 0.5 may represent any accuracy. For our reliability, only rel $= 0$ is similarly indescriptive, while all other values come with the following two guarantees for certainty *and* accuracy.

$$\text{cert} \geq |\text{rel}|, \qquad \text{acc}_{\text{calib}} \begin{cases} \geq \text{rel} & \text{for } \text{acc}_{\text{calib}} > 0 \\ \leq \text{rel} & \text{for } \text{acc}_{\text{calib}} < 0 \end{cases}. \tag{3}$$

A4. PARC's Calibration: Demonstration

Our explanation of PARC's calibration step in Fig. 2 was limited to a few selected methods. Therefore, we show the same evaluation but for all 22 evaluated methods in Fig. A2. In addition to the calibration results on single datasets, we further show that the expected trend – VLMs perform better on original than negated prompts – holds across all 6 balanced datasets, and is not limited to NYU-Depth V2 [16].

A5. Accompanying Results

Due to space limitation, we reported only a subset of PARC's metrics and potential evaluations in the main paper. Below, we show extended metrics for the most perturbing prompts as well as additional analysis regarding the most prompt-agnostic VLMs.

A5.1. To Which Prompts are VLMs Sensitive?

In Tab. A3 we show the additional detailed accuracy and certainty measures for our prompt type analysis in Tab. 3. The trends follow our reliability score.

Combined Vision-Language Variations. To test how combinations of prompt variations affect models, we select

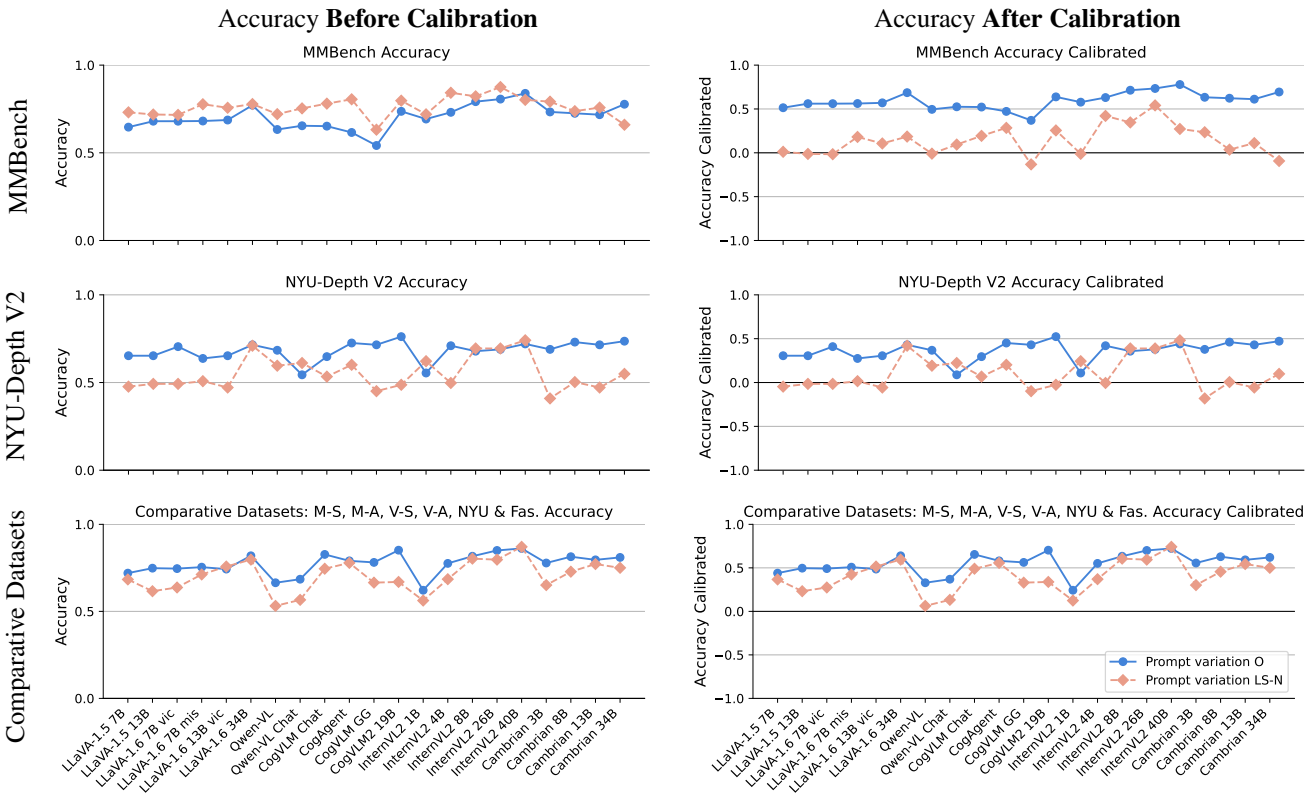

Figure A2. Effect of calibration, extension of Fig. 2 for all methods. Accuracy before *[Left]* vs. after calibration *[Right]* for original *[Blue]* and negated *[orange]* prompts. Only calibration aligns the ordering – original over negation – on the imbalanced MMBench [13] dataset with the balanced NYU-Depth V2 dataset [16] and the average over all balanced comparative datasets [9]. In comparison to Fig. 2, this plot shows results for all methods and also the averages over all six balanced datasets in the last row.

		Reliability Calibrated								Accuracy Calibrated								Certainty Calibrated								Consistency Calibrated							
Prompt Variation		M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**
Original	**O**	0.52	0.12	0.24	0.23	0.13	0.32	0.48	0.29	0.59	0.51	0.53	0.49	0.31	0.44	0.57	0.49	0.83	0.26	0.49	0.46	0.41	0.64	0.73	0.54								
Language Ref.	**LR-I**	0.51	0.07	0.22	0.16	0.13	0.29	0.47	0.26	0.60	0.41	0.46	0.44	0.30	0.43	0.57	0.46	0.78	0.17	0.45	0.37	0.41	0.57	0.72	0.50	0.83	0.68	0.74	0.73	0.79	0.80	0.87	0.78
	LR-C	0.41	0.13	0.17	0.15	0.11	0.28	0.45	0.24	0.53	0.46	0.44	0.46	0.27	0.39	0.56	0.45	0.69	0.27	0.42	0.35	0.40	0.56	0.70	0.49	0.78	0.73	0.71	0.79	0.64	0.63	0.84	0.73
	LR-V	0.35	0.22	0.14	0.12	0.09	0.09	0.42	0.21	0.54	0.57	0.34	0.32	0.27	0.34	0.53	0.42	0.58	0.35	0.44	0.34	0.32	0.31	0.69	0.43	0.72	0.69	0.63	0.62	0.49	0.49	0.79	0.63
Language Sem.	**LS-N**	0.39	0.06	0.22	0.18	0.04	0.25	0.05	0.17	0.56	0.52	0.47	0.44	0.13	0.45	0.14	0.39	0.61	0.11	0.41	0.35	0.23	0.48	0.38	0.37	0.58	0.50	0.42	0.44	0.06	0.37	0.06	0.35
	LS-A	0.30	0.14	0.11	0.10	0.00	0.05	-0.01	0.10	0.48	0.42	0.39	0.25	-0.03	0.24	-0.08	0.24	0.53	0.30	0.28	0.35	0.15	0.20	0.27	0.30	0.48	0.30	0.22	0.15	-0.34	0.01	-0.16	0.09
	LS-M	0.15	0.09	0.21	0.14	0.01	0.17	*n.a.*	0.13	0.39	0.57	0.47	0.38	-0.02	0.38	*n.a.*	0.36	0.35	0.15	0.44	0.31	0.16	0.40	*n.a.*	0.30	0.40	0.42	0.30	0.28	-0.26	0.28	*n.a.*	0.24
Vision Ref.	**VR-B**	0.41	0.09	0.23	0.22	0.13	0.32	0.43	0.26	0.49	0.43	0.49	0.50	0.31	0.45	0.54	0.46	0.76	0.25	0.47	0.43	0.40	0.63	0.69	0.52	0.75	0.71	0.80	0.80	0.77	0.83	0.84	0.79
	VR-R	0.48	0.21	0.23	0.29	0.08	0.13	0.38	0.26	0.58	0.47	0.49	0.50	0.26	0.28	0.49	0.44	0.79	0.42	0.48	0.57	0.35	0.38	0.65	0.52	0.79	0.70	0.68	0.72	0.53	0.56	0.74	0.67
	VR-L	0.33	0.12	0.15	0.20	0.09	0.19	0.41	0.21	0.43	0.42	0.34	0.46	0.26	0.33	0.52	0.39	0.65	0.25	0.39	0.41	0.33	0.41	0.67	0.45	0.56	0.63	0.55	0.61	0.62	0.61	0.79	0.63
Vision Sem.	**VS-S**	0.59	0.12	0.21	0.24	0.17	0.40	*n.a.*	0.29	0.68	0.50	0.48	0.54	0.39	0.57	*n.a.*	0.53	0.83	0.25	0.38	0.43	0.43	0.65	*n.a.*	0.49	0.65	0.56	0.42	0.46	0.23	0.45	*n.a.*	0.46
	VS-E	0.13	0.19	0.04	0.11	0.16	0.16	*n.a.*	0.13	0.28	0.41	0.30	0.31	0.38	0.30	*n.a.*	0.33	0.41	0.45	0.15	0.34	0.42	0.39	*n.a.*	0.36	0.13	0.09	-0.04	-0.04	-0.14	-0.02	*n.a.*	-0.00

Table A3. Most perturbing prompt variations in language and vision, averaged across models – additional measurements for Tab. 2. All scores are calibrated, with 1.0 being ideal model performance and 0.0 is random performance. The consistency is calculated between each varied and the original prompt. Most perturbing prompt per variation class is bold, and across variations underlined. High number indicates model robustness to a prompt variant. Prompt variation acronyms are from Tab. 1. MMBench's non-comparative questions about single images neither support more-less reformulation (LS-M), nor image swaps or exchanges (VS-S, VS-E).

one variation per type (LR-I, LS-N, VR-L, VS-S) and evaluate the variation combinations LR-I/VR-L, LR-I/VS-S, LS-N/VR-L and LS-N/VS-S, resulting in 28 additional datasets per VLM. In Tab. A6, we find that variation combinations behave like the worst variation or worse. As combined variations follow the individual results, PARC's per-variation evaluation is a cost-effective strategy to avoid combination explosions.

A5.2. Which VLM is Most Prompt Agnostic?

Below, we describe additional evaluation analysis for VLM prompt sensitivity. The first set of result is centered around expanding on Tab. 3. Then, we expand our arguments why data is most important for VLM prompt sensitivity.

VLM sensitivity across datasets. We also report detailed accuracy and certainty measures for our VLM sensitivity analysis in Tab. A5, expanding results from Tab. 3. Again, accuracy and certainty trends follow our reliability score.

	Reliability Calibrated								Accuracy Calibrated								Certainty Calibrated								Consistency Calibrated							
Model	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**
LLaVA-1.5 7B	0.33	0.09	0.12	0.12	0.06	0.12	0.28	0.16	0.57	0.32	0.33	0.38	0.16	0.28	0.39	0.35	0.56	0.29	0.34	0.29	0.29	0.35	0.56	0.38	0.62	0.40	0.48	0.52	0.29	0.42	0.60	0.48
LLaVA-1.5 13B	0.27	0.17	0.17	0.14	0.08	0.11	0.31	0.18	0.46	0.47	0.37	0.42	0.18	0.35	0.40	0.38	0.55	0.35	0.44	0.31	0.33	0.31	0.58	0.41	0.62	0.56	0.52	0.53	0.23	0.50	0.60	0.51
LLaVA-1.6 7B vic	0.34	0.09	0.15	0.16	0.12	0.11	0.30	0.18	0.52	0.35	0.34	0.46	0.26	0.24	0.40	0.37	0.62	0.25	0.38	0.31	0.32	0.38	0.58	0.41	0.54	0.34	0.41	0.53	0.36	0.14	0.60	0.42
LLaVA-1.6 7B mis	0.36	0.08	0.20	0.28	0.06	0.19	0.35	0.22	0.55	0.39	0.44	0.58	0.19	0.41	0.46	0.43	0.65	0.20	0.46	0.46	0.28	0.42	0.63	0.44	0.56	0.56	0.48	0.62	0.31	0.59	0.62	0.54
LLaVA-1.6 13B vic	0.49	0.12	0.23	0.26	0.09	0.21	0.34	0.25	0.63	0.54	0.47	0.57	0.22	0.42	0.43	0.47	0.76	0.22	0.44	0.45	0.33	0.39	0.63	0.46	0.76	0.57	0.57	0.65	0.30	0.54	0.62	0.57
LLaVA-1.6 34B	0.53	0.25	0.32	0.36	0.15	0.25	0.44	0.33	0.67	0.67	0.55	0.65	0.35	0.39	0.49	0.54	0.78	0.39	0.55	0.55	0.40	0.60	0.70	0.57	0.79	0.69	0.60	0.68	0.41	0.61	0.62	0.63
Qwen-VL	0.03	0.01	0.03	0.06	0.04	0.04	0.21	0.06	0.06	0.07	0.16	0.26	0.13	0.12	0.26	0.15	0.33	0.18	0.25	0.25	0.24	0.28	0.51	0.29	0.00	0.14	0.11	0.24	0.16	-0.06	0.42	0.14
Qwen-VL Chat	0.11	0.05	0.09	0.15	0.04	0.05	0.25	0.11	0.29	0.35	0.30	0.34	0.14	0.16	0.35	0.27	0.35	0.15	0.29	0.32	0.26	0.31	0.53	0.32	0.31	0.40	0.40	0.42	0.35	0.21	0.52	0.37
CogVLM Chat	0.48	0.12	0.32	0.22	0.09	0.34	0.26	0.26	0.67	0.58	0.61	0.54	0.20	0.47	0.36	0.49	0.68	0.20	0.49	0.40	0.34	0.62	0.55	0.47	0.71	0.63	0.67	0.60	0.30	0.50	0.62	0.58
CogAgent	0.42	0.12	0.21	0.14	0.10	0.32	0.23	0.22	0.58	0.54	0.56	0.47	0.26	0.47	0.34	0.46	0.71	0.22	0.36	0.31	0.34	0.58	0.55	0.44	0.70	0.60	0.59	0.56	0.40	0.52	0.64	0.57
CogVLM GG	-0.10	-0.07	-0.26	-0.23	-0.04	-0.15	0.02	-0.12	-0.24	-0.22	-0.16	-0.22	-0.01	-0.26	0.09	-0.15	0.53	0.35	0.44	0.37	0.38	0.35	0.39	0.40	0.06	-0.01	-0.06	-0.12	-0.10	-0.06	0.45	0.02
CogVLM2 19B	0.49	0.11	0.20	0.20	0.16	0.42	0.26	0.26	0.64	0.65	0.53	0.39	0.29	0.56	0.36	0.49	0.71	0.18	0.38	0.48	0.45	0.69	0.52	0.49	0.68	0.68	0.51	0.38	0.30	0.58	0.39	0.50
InternVL2 1B	0.13	0.03	0.01	0.04	0.05	0.08	0.30	0.09	0.29	0.20	0.09	0.13	0.14	0.18	0.35	0.20	0.36	0.16	0.18	0.28	0.25	0.21	0.53	0.28	0.32	0.40	0.31	0.18	0.15	0.26	0.54	0.31
InternVL2 2B	0.23	0.15	0.11	0.11	0.09	0.14	0.37	0.17	0.39	0.49	0.44	0.33	0.25	0.31	0.47	0.38	0.49	0.29	0.22	0.32	0.28	0.39	0.63	0.38	0.48	0.53	0.44	0.42	0.29	0.40	0.64	0.46
InternVL2 4B	0.35	0.21	0.23	0.18	0.13	0.29	0.41	0.26	0.48	0.60	0.50	0.43	0.29	0.51	0.52	0.47	0.69	0.36	0.44	0.39	0.36	0.53	0.67	0.49	0.67	0.61	0.50	0.59	0.34	0.52	0.64	0.55
InternVL2 8B	0.51	0.23	0.30	0.24	0.11	0.34	0.48	0.32	0.65	0.68	0.62	0.46	0.34	0.55	0.59	0.56	0.74	0.36	0.45	0.50	0.31	0.57	0.70	0.52	0.73	0.72	0.61	0.61	0.37	0.70	0.70	0.63
InternVL2 26B	0.67	0.21	0.31	0.30	0.16	0.50	0.52	0.38	0.75	0.74	0.64	0.55	0.36	0.62	0.63	0.61	0.86	0.28	0.46	0.55	0.42	0.74	0.73	0.58	0.83	0.77	0.69	0.70	0.47	0.69	0.72	0.70
InternVL2 40B	0.69	0.20	0.33	0.33	0.17	0.48	0.57	0.40	0.77	0.73	0.78	0.59	0.38	0.65	0.65	0.65	0.85	0.27	0.42	0.56	0.40	0.71	0.77	0.57	0.86	0.81	0.74	0.67	0.43	0.72	0.72	0.71
Cambrian 3B	0.48	0.16	0.18	0.11	0.10	0.23	0.39	0.24	0.61	0.53	0.42	0.37	0.22	0.51	0.49	0.45	0.75	0.31	0.39	0.26	0.34	0.43	0.66	0.45	0.72	0.55	0.57	0.50	0.28	0.57	0.62	0.54
Cambrian 8B	0.52	0.16	0.24	0.28	0.12	0.26	0.40	0.28	0.67	0.54	0.50	0.56	0.29	0.51	0.46	0.51	0.74	0.29	0.45	0.46	0.35	0.47	0.66	0.49	0.80	0.69	0.54	0.57	0.43	0.57	0.60	0.60
Cambrian 13B	0.51	0.20	0.23	0.20	0.11	0.25	0.38	0.27	0.64	0.61	0.53	0.48	0.25	0.47	0.46	0.49	0.77	0.33	0.44	0.41	0.34	0.48	0.65	0.49	0.76	0.62	0.62	0.53	0.29	0.53	0.61	0.57
Cambrian 34B	0.55	0.18	0.28	0.25	0.13	0.27	0.46	0.30	0.65	0.63	0.52	0.61	0.30	0.51	0.45	0.52	0.84	0.28	0.53	0.42	0.37	0.51	0.71	0.52	0.84	0.77	0.59	0.71	0.41	0.59	0.60	0.64

Table A4. Evaluation of model prompt sensitivity. Reliability, accuracy, certainty and consistency measured across models and datasets, and averaged over prompt variants. This expands Tab. 3 with the detailed numbers for accuracy and certainty. High numbers indicate high robustness. Note that CogVLM GG shows bad performance across all metrics, unless robustness, accuracy and consistency are calculated with the hightest-ranking logit score across answer choices in Tab. A7 rather than the direct VLM output.

	Reliability Calibrated								Accuracy Calibrated								Certainty Calibrated								Consistency Calibrated							
Model	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**
LLaVA-1.5 7B	0.33	0.09	0.12	0.12	0.06	0.12	0.28	0.16	0.57	0.32	0.33	0.38	0.16	0.28	0.39	0.35	0.56	0.29	0.34	0.29	0.29	0.35	0.56	0.38	0.62	0.40	0.48	0.52	0.29	0.42	0.60	0.48
LLaVA-1.5 13B	0.27	0.17	0.17	0.14	0.08	0.11	0.31	0.18	0.46	0.47	0.37	0.42	0.18	0.35	0.40	0.38	0.55	0.35	0.44	0.31	0.33	0.31	0.58	0.41	0.62	0.56	0.52	0.53	0.23	0.50	0.60	0.51
LLaVA-1.6 7B vic	0.34	0.09	0.15	0.16	0.12	0.11	0.30	0.18	0.52	0.35	0.34	0.46	0.26	0.24	0.40	0.37	0.62	0.25	0.38	0.31	0.32	0.38	0.58	0.41	0.54	0.34	0.41	0.53	0.36	0.14	0.60	0.42
LLaVA-1.6 7B mis	0.36	0.08	0.20	0.28	0.06	0.19	0.35	0.22	0.55	0.39	0.44	0.58	0.19	0.41	0.46	0.43	0.65	0.20	0.46	0.46	0.28	0.42	0.63	0.44	0.56	0.56	0.48	0.62	0.31	0.59	0.62	0.54
LLaVA-1.6 13B vic	0.49	0.12	0.23	0.26	0.09	0.21	0.34	0.25	0.63	0.54	0.47	0.57	0.22	0.42	0.43	0.47	0.76	0.22	0.44	0.45	0.33	0.39	0.63	0.46	0.76	0.57	0.57	0.65	0.30	0.54	0.62	0.57
LLaVA-1.6 34B	0.53	0.25	0.32	0.36	0.15	0.25	0.44	0.33	0.67	0.67	0.55	0.65	0.35	0.39	0.49	0.54	0.78	0.39	0.55	0.55	0.40	0.60	0.70	0.57	0.79	0.69	0.60	0.68	0.41	0.61	0.62	0.63
Qwen-VL	0.03	0.01	0.03	0.06	0.04	0.04	0.21	0.06	0.06	0.07	0.16	0.26	0.13	0.12	0.26	0.15	0.33	0.18	0.25	0.25	0.24	0.28	0.51	0.29	0.00	0.14	0.11	0.24	0.16	-0.06	0.42	0.14
Qwen-VL Chat	0.11	0.05	0.09	0.15	0.04	0.05	0.25	0.11	0.29	0.35	0.30	0.34	0.14	0.16	0.35	0.27	0.35	0.15	0.29	0.32	0.26	0.31	0.53	0.32	0.31	0.40	0.40	0.42	0.35	0.21	0.52	0.37
CogVLM Chat	0.48	0.12	0.32	0.22	0.09	0.34	0.26	0.26	0.67	0.58	0.61	0.54	0.20	0.47	0.36	0.49	0.68	0.20	0.49	0.40	0.34	0.62	0.55	0.47	0.71	0.63	0.67	0.60	0.30	0.50	0.62	0.58
CogAgent	0.42	0.12	0.21	0.14	0.10	0.32	0.23	0.22	0.58	0.54	0.56	0.47	0.26	0.47	0.34	0.46	0.71	0.22	0.36	0.31	0.34	0.58	0.55	0.44	0.70	0.60	0.59	0.56	0.40	0.52	0.64	0.57
CogVLM GG	-0.10	-0.07	-0.26	-0.23	-0.04	-0.15	0.02	-0.12	-0.24	-0.22	-0.16	-0.22	-0.01	-0.26	0.09	-0.15	0.53	0.35	0.44	0.37	0.38	0.35	0.39	0.40	0.06	-0.01	-0.06	-0.12	-0.10	-0.06	0.45	0.02
CogVLM2 19B	0.49	0.11	0.20	0.20	0.16	0.42	0.26	0.26	0.64	0.65	0.53	0.39	0.29	0.56	0.36	0.49	0.71	0.18	0.38	0.48	0.45	0.69	0.52	0.49	0.68	0.68	0.51	0.38	0.30	0.58	0.39	0.50
InternVL2 1B	0.13	0.03	0.01	0.04	0.05	0.08	0.30	0.09	0.29	0.20	0.09	0.13	0.14	0.18	0.35	0.20	0.36	0.16	0.18	0.28	0.25	0.21	0.53	0.28	0.32	0.40	0.31	0.18	0.15	0.26	0.54	0.31
InternVL2 2B	0.23	0.15	0.11	0.11	0.09	0.14	0.37	0.17	0.39	0.49	0.44	0.33	0.25	0.31	0.47	0.38	0.49	0.29	0.22	0.32	0.28	0.39	0.63	0.38	0.48	0.53	0.44	0.42	0.29	0.40	0.64	0.46
InternVL2 4B	0.35	0.21	0.23	0.18	0.13	0.29	0.41	0.26	0.48	0.60	0.50	0.43	0.29	0.51	0.52	0.47	0.69	0.36	0.44	0.39	0.36	0.53	0.67	0.49	0.67	0.61	0.50	0.59	0.34	0.52	0.64	0.55
InternVL2 8B	0.51	0.23	0.30	0.24	0.11	0.34	0.48	0.32	0.65	0.68	0.62	0.46	0.34	0.55	0.59	0.56	0.74	0.36	0.45	0.50	0.31	0.57	0.70	0.52	0.73	0.72	0.61	0.61	0.37	0.70	0.70	0.63
InternVL2 26B	0.67	0.21	0.31	0.30	0.16	0.50	0.52	0.38	0.75	0.74	0.64	0.55	0.36	0.62	0.63	0.61	0.86	0.28	0.46	0.55	0.42	0.74	0.73	0.58	0.83	0.77	0.69	0.70	0.47	0.69	0.72	0.70
InternVL2 40B	0.69	0.20	0.33	0.33	0.17	0.48	0.57	0.40	0.77	0.73	0.78	0.59	0.38	0.65	0.65	0.65	0.85	0.27	0.42	0.56	0.40	0.71	0.77	0.57	0.86	0.81	0.74	0.67	0.43	0.72	0.72	0.71
Cambrian 3B	0.48	0.16	0.18	0.11	0.10	0.23	0.39	0.24	0.61	0.53	0.42	0.37	0.22	0.51	0.49	0.45	0.75	0.31	0.39	0.26	0.34	0.43	0.66	0.45	0.72	0.55	0.57	0.50	0.28	0.57	0.62	0.54
Cambrian 8B	0.52	0.16	0.24	0.28	0.12	0.26	0.40	0.28	0.67	0.54	0.50	0.56	0.29	0.51	0.46	0.51	0.74	0.29	0.45	0.46	0.35	0.47	0.66	0.49	0.80	0.69	0.54	0.57	0.43	0.57	0.60	0.60
Cambrian 13B	0.51	0.20	0.23	0.20	0.11	0.25	0.38	0.27	0.64	0.61	0.53	0.48	0.25	0.47	0.46	0.49	0.77	0.33	0.44	0.41	0.34	0.48	0.65	0.49	0.76	0.62	0.62	0.53	0.29	0.53	0.61	0.57
Cambrian 34B	0.55	0.18	0.28	0.25	0.13	0.27	0.46	0.30	0.65	0.63	0.52	0.61	0.30	0.51	0.45	0.52	0.84	0.28	0.53	0.42	0.37	0.51	0.71	0.52	0.84	0.77	0.59	0.71	0.41	0.59	0.60	0.64

Table A5. Evaluation of model prompt sensitivity. Reliability, accuracy, certainty and consistency measured across models and datasets, and averaged over prompt variants. This expands Tab. 3 with the detailed numbers for accuracy and certainty. High numbers indicate high robustness. Note that CogVLM GG shows bad performance across all metrics, unless robustness, accuracy and consistency are calculated with the hightest-ranking logit score across answer choices in Tab. A7 rather than the direct VLM output.

	Rel	Acc	Cert	Con		Rel	Acc	Cert	Con
LR	0.26	0.46	0.50	0.78	**LS**	0.17	0.39	0.37	0.35
VR	0.21	0.39	0.45	0.63	**VR**	0.21	0.39	0.45	0.63
LR+VR	0.18	0.37	0.41	0.59	**LS+VR**	0.11	0.31	0.32	0.26
LR	0.26	0.46	0.50	0.78	**LS**	0.17	0.39	0.37	0.35
VS	0.29	0.53	0.49	0.46	**VS**	0.29	0.53	0.49	0.46
LR+VS	0.24	0.46	0.45	0.42	**LS+VS**	0.16	0.35	0.38	0.42

Table A6. Effect of combining vision and language prompt variations on reliability, accuracy, certainty, and consistency, when averaged across all datasets. Expands Tab. 2 and Tab. A3.

VLM sensitivity via logit evaluation. We can also evaluate VLMs by extracting the logit values for all potential multiple choice answers, *i.e.* the logits for `A`, `B`, `C` and `D`, and treat the highest as model answer. This simplifies parsing the VLMs response and matching it to the possible answers, but may also generate false positive answers if the VLM generates a free form answer that does not match the answer choices at all. We report the results in Tab. A7, which replicates Tab. 3 of the main paper with its extended results in Tab. A5. Overall, the trends are well aligned. The only major difference is CogVLM GG, which performs in line with the other CogVLM variants and does not underperform any longer if evaluated by logits. This indicates that CogVLM GG generates text that does not resemble any answer choice but would find the correct answer if the response outputs were limited to the answer options.

VLM sensitivity across variation classes. To complement the right figure in Tab. 3 of the main paper, we show the

Model	Reliability Calibrated								Accuracy Calibrated								Certainty Calibrated								Consistency Calibrated							
	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**	M-S	M-A	V-S	V-A	NYU	Fas	MMB	**AVG**
LLaVA-1.5 7B	0.33	0.09	0.12	0.12	0.06	0.12	0.28	0.16	0.57	0.32	0.33	0.38	0.16	0.28	0.39	0.35	0.56	0.29	0.34	0.29	0.29	0.35	0.56	0.38	0.62	0.40	0.48	0.52	0.29	0.42	0.60	0.48
LLaVA-1.5 13B	0.27	0.17	0.17	0.14	0.08	0.11	0.31	0.18	0.46	0.47	0.37	0.42	0.18	0.35	0.42	0.38	0.55	0.35	0.44	0.31	0.33	0.31	0.58	0.41	0.62	0.56	0.52	0.53	0.23	0.50	0.60	0.51
LLaVA-1.6 7B vic	0.34	0.09	0.15	0.16	0.12	0.11	0.30	0.18	0.52	0.35	0.34	0.46	0.26	0.24	0.41	0.37	0.62	0.25	0.38	0.31	0.32	0.38	0.58	0.41	0.54	0.34	0.41	0.53	0.36	0.14	0.59	0.42
LLaVA-1.6 7B mis	0.36	0.08	0.20	0.28	0.06	0.19	0.35	0.22	0.55	0.39	0.44	0.58	0.19	0.41	0.46	0.43	0.65	0.20	0.46	0.46	0.28	0.42	0.63	0.44	0.56	0.56	0.48	0.62	0.31	0.59	0.62	0.54
LLaVA-1.6 13B vic	0.49	0.12	0.23	0.26	0.09	0.21	0.34	0.25	0.63	0.54	0.47	0.57	0.22	0.42	0.44	0.47	0.76	0.22	0.44	0.45	0.33	0.39	0.63	0.46	0.76	0.57	0.57	0.65	0.30	0.54	0.62	0.57
LLaVA-1.6 34B	0.53	0.25	0.33	0.36	0.15	0.27	0.46	0.33	0.66	0.65	0.56	0.65	0.37	0.43	0.54	0.55	0.78	0.39	0.55	0.55	0.40	0.60	0.70	0.57	0.81	0.70	0.67	0.67	0.50	0.66	0.65	0.67
Qwen-VL	0.15	0.04	0.07	0.07	0.06	0.07	0.23	0.10	0.38	0.18	0.28	0.27	0.22	0.21	0.34	0.27	0.33	0.18	0.25	0.25	0.24	0.28	0.51	0.29	0.23	0.31	0.29	0.31	0.36	0.09	0.53	0.30
Qwen-VL Chat	0.09	0.07	0.03	0.12	0.02	0.07	0.25	0.09	0.24	0.40	0.16	0.26	0.09	0.19	0.38	0.24	0.35	0.15	0.29	0.32	0.26	0.31	0.53	0.32	0.24	0.39	0.11	0.23	0.19	0.19	0.51	0.27
CogVLM Chat	0.48	0.12	0.32	0.22	0.09	0.34	0.28	0.26	0.67	0.58	0.62	0.54	0.20	0.47	0.41	0.50	0.68	0.20	0.49	0.40	0.34	0.62	0.55	0.47	0.71	0.63	0.67	0.61	0.30	0.50	0.62	0.58
CogAgent	0.42	0.12	0.21	0.14	0.10	0.32	0.26	0.22	0.58	0.54	0.56	0.47	0.26	0.47	0.40	0.47	0.71	0.22	0.36	0.31	0.34	0.58	0.55	0.44	0.70	0.60	0.59	0.56	0.40	0.52	0.64	0.57
CogVLM GG	0.29	0.14	0.21	0.19	0.11	0.13	0.11	0.17	0.47	0.38	0.35	0.42	0.23	0.34	0.25	0.35	0.53	0.35	0.44	0.37	0.38	0.35	0.39	0.40	0.55	0.45	0.33	0.43	0.26	0.35	0.50	0.41
CogVLM2 19B	0.54	0.11	0.24	0.24	0.18	0.47	0.26	0.29	0.72	0.67	0.64	0.45	0.33	0.62	0.37	0.54	0.71	0.18	0.38	0.48	0.45	0.69	0.52	0.49	0.79	0.71	0.57	0.53	0.37	0.69	0.47	0.59
InternVL2 1B	0.13	0.03	0.01	0.04	0.05	0.08	0.30	0.09	0.29	0.20	0.09	0.13	0.14	0.18	0.40	0.20	0.36	0.16	0.18	0.28	0.25	0.21	0.53	0.28	0.32	0.40	0.31	0.18	0.15	0.26	0.54	0.31
InternVL2 2B	0.23	0.15	0.11	0.11	0.09	0.14	0.37	0.17	0.39	0.49	0.44	0.33	0.25	0.31	0.47	0.38	0.49	0.29	0.22	0.32	0.28	0.39	0.63	0.38	0.48	0.53	0.44	0.42	0.29	0.40	0.64	0.46
InternVL2 4B	0.35	0.21	0.21	0.19	0.12	0.27	0.41	0.25	0.47	0.59	0.46	0.40	0.27	0.47	0.52	0.46	0.69	0.36	0.44	0.39	0.36	0.53	0.67	0.49	0.64	0.56	0.52	0.54	0.34	0.46	0.65	0.53
InternVL2 8B	0.51	0.23	0.30	0.24	0.11	0.34	0.48	0.32	0.65	0.68	0.62	0.46	0.34	0.55	0.59	0.56	0.74	0.36	0.45	0.50	0.31	0.57	0.70	0.52	0.73	0.72	0.61	0.61	0.37	0.70	0.69	0.63
InternVL2 26B	0.67	0.21	0.31	0.30	0.16	0.50	0.52	0.38	0.75	0.74	0.64	0.55	0.36	0.62	0.62	0.61	0.86	0.28	0.46	0.55	0.42	0.74	0.73	0.58	0.83	0.77	0.69	0.70	0.47	0.69	0.72	0.70
InternVL2 40B	0.69	0.20	0.33	0.33	0.17	0.48	0.56	0.40	0.77	0.73	0.78	0.59	0.38	0.65	0.61	0.64	0.85	0.27	0.42	0.56	0.40	0.71	0.77	0.57	0.86	0.81	0.74	0.67	0.43	0.72	0.68	0.70
Cambrian 3B	0.46	0.16	0.17	0.11	0.10	0.23	0.39	0.23	0.59	0.51	0.40	0.38	0.21	0.50	0.49	0.44	0.75	0.31	0.39	0.26	0.34	0.43	0.66	0.45	0.70	0.51	0.54	0.51	0.26	0.54	0.64	0.53
Cambrian 8B	0.52	0.15	0.23	0.28	0.12	0.26	0.40	0.28	0.67	0.52	0.48	0.56	0.30	0.50	0.47	0.50	0.74	0.29	0.45	0.46	0.35	0.47	0.66	0.49	0.80	0.62	0.54	0.58	0.40	0.56	0.60	0.59
Cambrian 13B	0.50	0.20	0.24	0.20	0.12	0.24	0.38	0.27	0.62	0.60	0.54	0.48	0.25	0.46	0.46	0.49	0.77	0.33	0.44	0.41	0.34	0.48	0.65	0.49	0.78	0.66	0.62	0.48	0.29	0.52	0.61	0.57
Cambrian 34B	0.55	0.18	0.28	0.26	0.13	0.27	0.46	0.30	0.64	0.64	0.51	0.62	0.31	0.50	0.49	0.53	0.84	0.28	0.53	0.42	0.37	0.51	0.71	0.52	0.84	0.76	0.60	0.69	0.43	0.59	0.60	0.64

Table A7. Evaluation of model prompt sensitivity with hightest logit score across answer options instead of raw VLM output. This mirrors Tab. A5. Note that the certainty values are unaffected by this change because certainty scores via conformal predictions already use logit scores. The trends and model behaviors remain almost unchanged, except for CogVLM GG, which now follows the trend of other CogVLM variants instead of underperforming significantly.

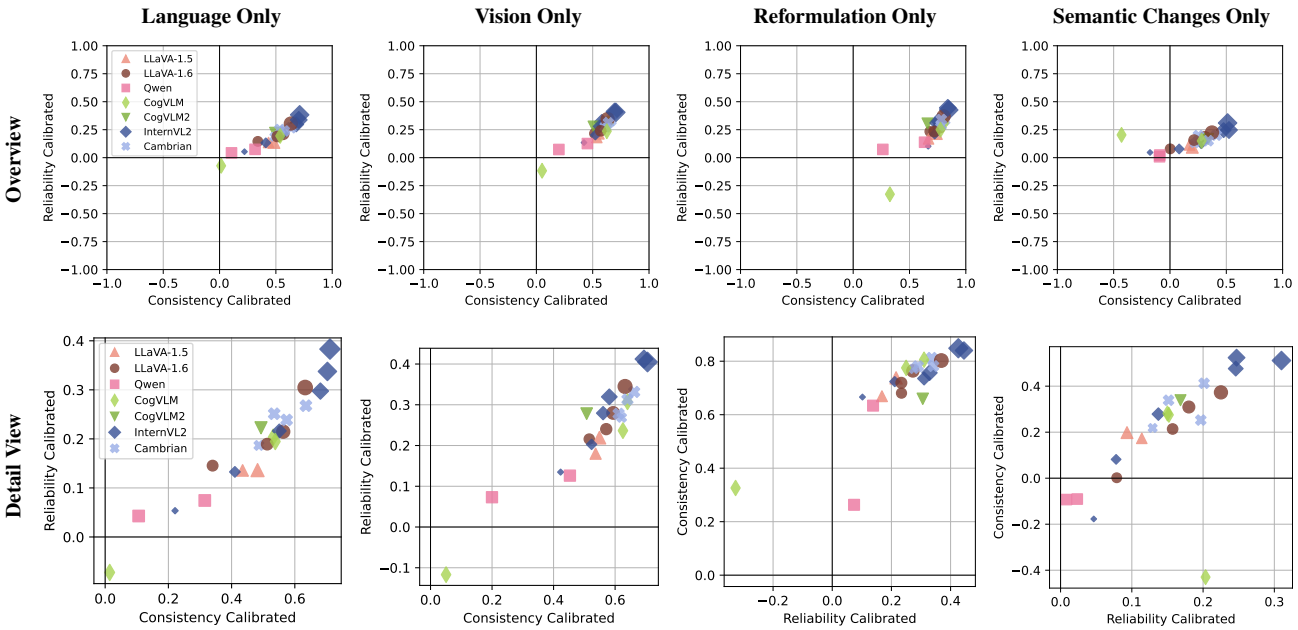

Figure A3. Consistency vs. reliability over the four prompt variation types: Language, Vision, Reformulations and Semantic Changes. More detailed analysis to the overall averages in Tab. 3. The top row shows the full ranges for calibrated reliability and consistency (from -1 to 1). The lower row shows zoom-ins. Marker size indicates VLM size. Model family rankings are stable across datasets. The most robust model family to prompt changes is InternVL2, with 40B as best.

consistency vs. reliability plots for our four main prompt variation types: Language variations, Vision variations, Reformulations and Semantic Changes in Fig. A3. Across all variations, the ranking of methods remain very similar, and model families continue to be the main influence in VLM performance. InternVL2 40B is still the most robust model, and within model families larger models also have an increased resilience against prompt variations. The detailed analysis further reinforces our findings on prompt variation types: Reformulations are easier for VLMs to understand, and excluding semantic changes from sensitivity evaluations gives the wrong appearance of an overall improved prompt-agnosticism.

VLM sensitivity across VLM and LLM sizes. To complement the analyses in Fig. 4, we show reliability and consistency measurements over model- and LLM size in Fig. A4.

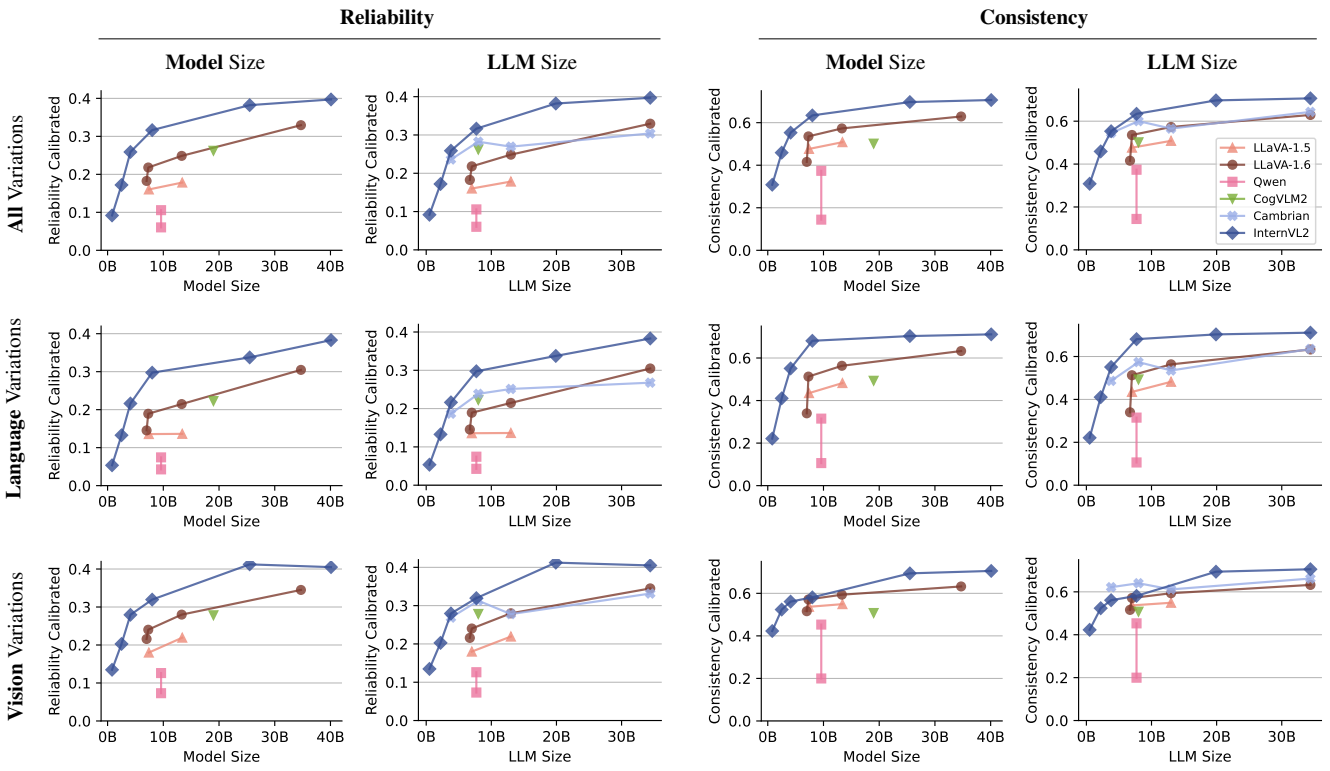

Figure A4. Influence of Model and VLM size on reliability and consistency, extension of Fig. 4. The first row shows averages across all prompt variations, while the second and third row show averages over only language and only vision variations. The trends do not change when only one prompt modality is varied. Also, the trends across reliability and consistency remain comparable.

In contrast to the main paper, we also include the reliability and consistency measurements over *only* language and *only* vision variations. Interestingly, the trends are very well aligned across language and vision variations, which further reinforces that architectural differences in the vision and language processing blocks do not sufficiently explain the differences in VLM prompt sensitivity.

References

[1] LLAMA-3: https://ai.meta.com/blog/meta-llama-3/. 1, 3

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 4

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 1

[5] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 1

[6] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 1

[7] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, 2015. 3

[8] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Proc. European Conference on Computer Vision (ECCV)*, pages 316–332. Springer, 2020. 3

[9] Jihyung Kil, Zheda Mai, Justin Lee, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Arpita Chowdhury, and Wei-Lun Chao. CompBench: A comparative reasoning benchmark for multimodal llms. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3, 5

[10] Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*, 2024. 4

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 1

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. https://llava-vl.github.io/blog/2024-01-30-llava-next/. 1

[13] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *Proc. European Conference on Computer Vision (ECCV)*, pages 216–233. Springer, 2025. 1, 3, 5

[14] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13018–13028, 2021. 3

[15] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 35:34405–34420, 2022. 1

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 3, 4, 5

[17] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024. 1

[18] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[19] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking LLMs via uncertainty quantification. *arXiv preprint arXiv:2401.12794*, 2024. 4