CONDITIONAL CONFIDENCE CALIBRATION FOR FOUNDATION MODEL DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models increasingly serve as decision components, where discrete outputs require reliable confidence estimates. In safety applications, vision-language and language models make fine-grained judgments about image or text content. These multi-label tasks decompose naturally into binary decisions, each requiring a calibrated confidence estimate. The same need arises whenever a model score is used to accept, reject, or route an output. Standard post-hoc calibration typically learns a global or category-level mapping from model scores to confidence, potentially missing systematic reliability differences across input contexts.

We propose *conditional confidence calibration*, which estimates the probability of a defined binary outcome from the model score, the decision being calibrated, and context available at decision time. Standard global and per-category calibration use only the score and, optionally, the decision category. Our adaptive method additionally fits a contextual calibrator using input descriptors such as text load, visual realism, or request domain, then learns from calibration data how much it should modify the better-supported decision-level estimate. The learned combination can default to that estimate when context is unhelpful, without updating the base foundation model.

We evaluate across seven image, image-text, and text safety datasets and three non-safety datasets using Gemma 4 models, with E4B throughout and matched 12B evaluations on two image-text datasets. Across 14 safety comparisons, adaptive calibration improves Macro balanced accuracy over per-category confidence in 12, with mean relative gains of 5.3% on image and image-text tasks and 5.7% on text tasks, and lowers Brier score by 6.3% on average. On the non-safety datasets, it lowers Brier score by 3.2–12.1% and improves balanced accuracy by 2.0–2.8%. The gains also hold across model families: with Qwen3-VL and InternVL3.5 at 4B and 8B, adaptive calibration improves both metrics in all eight comparisons. These results demonstrate that adaptive combination improves both decision and probability quality over non-contextual confidence baselines.

1 INTRODUCTION

Foundation models are increasingly used inside systems that make decisions, not only generate text. In those settings, a discrete output is incomplete without a reliable estimate of when the output should be trusted. Prior work shows that model probabilities and expressed confidence are useful but often miscalibrated (Guo et al., 2017; Desai & Durrett, 2020; Jiang et al., 2020; Kadavath et al., 2022; Lin et al., 2022). Better confidence estimates can guide deployment decisions and are also useful for core model-development work: prioritizing data collection, filtering noisy synthetic data, selecting examples for finetuning and alignment, comparing checkpoints, regression testing, and drift monitoring (Hendrycks & Gimpel, 2016; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Geifman & El-Yaniv, 2017; 2019).

Most post-hoc calibration learns a global or task-level mapping from model scores to probabilities, for example through Platt scaling, isotonic regression, Bayesian binning, or temperature scaling (Platt, 1999; Zadrozny & Elkan, 2002; Pakdaman Naeini et al., 2015; Guo et al., 2017). This can be too coarse even within one decision category. For example, among inputs evaluated for information safety, a score of 0.8 may correspond to different empirical positive rates for text-heavy screenshots

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Estimator	Conditions on	Can address	Main limitation
Global	Score only	Overall score bias	Category and context shifts
Decision-specific	Score and decision type	Decision-specific score bias	Within-decision context shifts
Attribute-guided	Score, decision type, and attributes	Supported context-group shifts	Sparse or noisy attributes
Adaptive	Score-only and contextual estimates	Context-dependent reliability shifts	Limited calibration support

Table 1: Confidence calibrators differ in the information they condition on. Score-only means using the model's answer-option score without decision identity or content context. Global and decision-specific calibration are standard special cases; adaptive calibration is the proposed estimator that learns when to use context. Section 4.3 describes the methods in detail.

and ordinary photographs. A per-category calibrator maps equal scores to equal probabilities and therefore cannot correct this within-category reliability shift. The same issue extends beyond safety to document review, medical advice, financial guidance, policy compliance, and visual quality control.

We propose *conditional confidence calibration*: reliability is estimated from model evidence, the binary decision being made, and context available at decision time. More precisely, confidence denotes the probability of a defined binary outcome. A binary decision is the unit of evaluation in this paper, such as whether an image contains violence or whether a text request asks for financial advice. This unit is natural for multi-label safety data, but the same setup applies whenever a system produces scalar evidence for a defined yes/no event. For example, a router can estimate the probability that a proposed route will succeed from its routing score and the request type available before routing. For safety classification, the outcome is label presence; for the non-safety tasks, it is whether the model's task prediction is correct.

The framework is useful because it turns calibration into a design question: what should the confidence estimate condition on? Standard global and per-category calibrators are recovered as special cases that condition on less information. More contextual calibrators can correct repeatable local errors, but they also need enough calibration support. For example, suppose a model's high confidence for an information-safety label is reliable on average but systematically overconfident on text-heavy screenshots. A supported text-load context can lower confidence for those inputs while retaining the better-supported category-level estimate elsewhere. Table 1 summarizes these alternatives: global and decision-specific estimates provide the coarser fallback, attribute-guided calibration handles supported content groups, and adaptive calibration learns when to apply the contextual estimate.

We use attribute-guided calibration as an interpretable contextual instantiation and propose adaptive calibration as our main method. Attribute-guided calibration uses decision-time descriptors such as visual realism, text load, subject family, scene privacy, request domain, or interaction framing. These descriptors characterize input context rather than the target label and can be observed directly or extracted by an auxiliary model. Adaptive calibration combines a better-supported decision-level estimate with a more specific context-conditioned estimate. The same principle can combine more than two estimators; the two-estimator version makes the support tradeoff explicit and easy to test. We evaluate five policies: the model's discrete output, one global threshold, decision-specific thresholds, attribute-guided thresholds for supported groups, and adaptive calibration, which learns how strongly context should modify the decision-level estimate (Section 4.3).

We evaluate whether more specific context improves the objective used to select the calibrator. Macro balanced accuracy is primary because it weights positive and negative decisions equally within each imbalanced category. Brier score (mean squared probability error) is the primary probability metric because it is a proper scoring rule that avoids arbitrary confidence bins (Brier, 1950). Section 3 defines the remaining metrics and motivates our choice of primary and secondary diagnostics. Experiments cover image, image-text, and text safety decisions, three model families at multiple sizes, and three non-safety evaluations (Appendix A.4).

Our contributions are:

- We introduce conditional confidence calibration, which unifies global, decision-specific, and contextual estimators through their conditioning information and exposes a support–specificity tradeoff.
- We propose adaptive calibration, which learns how to combine decision-level and contextual calibrators and retains the former when context does not improve probability quality.
- Across image, image-text, and text safety evaluations, adaptive calibration improves Macro BA over per-category confidence in seven of eight settings and 12 of 14 matched comparisons, and lowers Brier score by 6.3% on average. Matched model-size, supported-group, and three non-safety datasets provide additional evidence for the adaptive mechanism and its scope.

2 Related Work

Post-hoc calibration. Classical post-hoc calibration maps model scores to probabilities using methods such as Platt scaling, isotonic regression, Bayesian binning, and temperature scaling (Platt, 1999; Zadrozny & Elkan, 2002; Pakdaman Naeini et al., 2015; Guo et al., 2017). Input-dependent temperature methods increase flexibility by learning prediction- or sample-specific mappings (Tomani et al., 2022; Joy et al., 2023). These approaches enrich the mapping from model evidence, whereas our framework makes the conditioning variables explicit and combines semantic decision-time context with a better-supported decision-specific fallback.

Group-aware calibration. Group-aware calibration is the closest statistical precedent. Multicalibration and multiaccuracy target systematic errors across intersecting subpopulations, while Mondrian conformal methods partition calibration data by known groups (Hébert-Johnson et al., 2018; Kim et al., 2018; Shafer & Vovk, 2008). Our contribution is not the general idea of conditional calibration, but its formulation for prompted foundation-model decisions using pre-decision content descriptors and an adaptive estimator that learns when a local contextual estimate should modify a coarser estimate.

Foundation-model uncertainty. LLM uncertainty work obtains reliability signals from token probabilities, self-evaluation, verbalized confidence, or black-box estimates (Jiang et al., 2020; Kadavath et al., 2022; Lin et al., 2022; Feng et al., 2025). Related work also studies training-time calibration, open-ended sequence likelihoods, selective prediction, and routing (Yaldiz et al., 2026; Zhao et al., 2022; Geifman & El-Yaniv, 2017; 2019). We instead keep the base model fixed and ask whether decision type and content context improve a post-hoc confidence estimator. Appendix A.1 provides fuller positioning.

3 Problem Setup

We study examples x_i with binary decisions $d \in \mathcal{D}(x_i)$. A decision can be a category in a multi-label task, a binary property queried during generation, or one of several decision points in a generated sequence. Each decision has a binary outcome $y_{i,d} \in \{0, 1\}$ and a model-derived score $s_{i,d} \in [0, 1]$. The score may be an answer-option probability, a classifier probability, a margin-derived score, or another scalar confidence signal. The outcome represents label presence in the safety experiments and prediction correctness in the non-safety tasks. The calibration problem is to estimate its probability under the available evidence and context, not to prescribe a single score-extraction interface. The method therefore requires access to a decision-level score; APIs that expose only final discrete outputs cannot be calibrated directly.

A decision policy maps score, decision type, and context to a binary output, $\hat{y}_{i,d} = h(s_{i,d}, d, Z_i)$, where Z_i is available before the decision is scored. Calibration parameters and policy choices are selected only on calibration data, then frozen before held-out evaluation. This protocol keeps context-conditioned rules post hoc and prevents test-set subset discovery from influencing policy selection. The selection objective is part of the problem: a policy selected for Macro balanced accuracy (Macro BA) or F1 need not be the same policy selected for expected calibration error (ECE). Our method instantiates this setup with a decision-level estimate $\hat{p}^{\text{score}} = g(s, d)$ and a contextual estimate $\hat{p}^{\text{ctx}} = g(s, d, Z)$. Adaptive calibration learns from calibration data how to combine them; Section 4.4 defines the estimator.

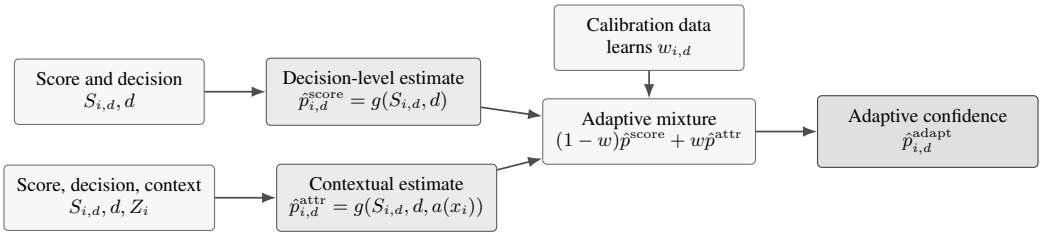


Figure 1: Adaptive calibration combines a decision-level confidence estimate with a more contextual estimate. The learned weight decides, for each example–decision pair, how much the contextual estimate should change the better-supported decision-level estimate.

For multi-label decision quality, the primary metric is Macro BA, which averages per-category balanced accuracy so positive and negative decisions receive equal weight inside each category. For decision type d, $\mathrm{BA}_d = \frac{1}{2}(\mathrm{TPR}_d + \mathrm{TNR}_d)$, and $\text{Macro BA} = |\mathcal{D}|^{-1} \sum_{d \in \mathcal{D}} \mathrm{BA}_d$. For probability quality, the primary metric is Brier score, computed on continuous probabilities for the binary decisions. Over N held-out example–decision pairs, $\text{Brier} = N^{-1} \sum_{i,d} (\hat{p}_{i,d} - y_{i,d})^2$, where $\hat{p}_{i,d}$ is the calibrated probability. Brier is a proper scoring rule and evaluates each probability against its binary outcome, so it captures both reliability and useful separation among examples (Brier, 1950; Murphy, 1973). We also report Macro F1, Micro F1, ECE, exact full-label-set accuracy, Hamming accuracy, and sample-level Jaccard as diagnostics. ECE bins held-out decision probabilities and averages the gap between empirical outcome frequency and mean probability in each bin; this makes it useful for reliability visualization, but sensitive to binning choices and possible cancellation within bins (Nixon et al., 2019).

4 Method

4.1 Conditional Confidence Calibration

For example i and binary decision d, let $S_{i,d}$ be raw model evidence, $Y_{i,d} \in \{0, 1\}$ the binary outcome being calibrated, and Z_i context available when that decision is scored. Conditional confidence calibration estimates

$$\hat{p}_{i,d} = g_\theta(S_{i,d}, d, Z_i) \approx \Pr(Y_{i,d} = 1 \mid S_{i,d}, d, Z_i). \tag{1}$$

This single form contains common post-hoc calibrators as special cases. A global calibrator uses $g(S)$, a decision-specific calibrator uses $g(S, d)$, and an attribute-conditioned calibrator uses $g(S, d, a(x))$, where $a(x)$ is a descriptor available at decision time. Table 1 summarizes the conditioning choices. The central statistical tradeoff is support: more context can reduce systematic calibration error when reliability varies across observable inputs, but small context groups require fallback to coarser estimates.

4.2 Evidence for Binary Decisions

The framework applies whenever the model provides scalar evidence for a binary decision. For prompted LLMs and VLMs, we form positive and negative answer-option sets $\mathcal{T}_d^+$ and $\mathcal{T}_d^-$ for decision d, then compute

$$s_{i,d} = \frac{\sum_{t \in \mathcal{T}_d^+} p(t \mid x_i, d)}{\sum_{t \in \mathcal{T}_d^+} p(t \mid x_i, d) + \sum_{t \in \mathcal{T}_d^-} p(t \mid x_i, d)}. \tag{2}$$

Multiple decision points in a generated sequence can each define their own binary decision d. Non-generative classifiers fit the same framework by replacing $s_{i,d}$ with a probability, margin-derived score, or sequence-level decision score.

4.3 Policy Family

We compare five policies. *Discrete generation* uses the generated binary decision directly. *Global thresholding* applies one threshold τ to the shared score-only probabilities for all decisions. *Decision-*

specific thresholding learns one threshold τ_d per decision type; in our multi-label safety experiments this is per-category thresholding. *Attribute-guided thresholding* starts from the decision-specific policy and uses a local decision-attribute threshold only when calibration support and selection performance justify it. *Adaptive calibration* first produces a calibrated probability for each example–decision pair by combining a decision-level estimate with a context-conditioned estimate. A decision threshold is then selected on calibration rows, exactly as for the other threshold policies, and applied unchanged to held-out rows.

Attributes instantiate Z_i in an interpretable form. They may come from dataset metadata, deterministic parsing, or model-assisted annotation, but they must not contain true labels, generated labels, correctness indicators, model scores, selected parameters, or evaluation metrics. This constraint keeps the problem a calibration problem rather than a leakage-driven label shortcut.

4.4 Adaptive Calibration

Our main contribution is an adaptive probability estimator that combines confidence estimates at different levels of contextual detail. In our experiments, the score-only estimate $\hat{p}^{\text{score}}_{i,d}$ is calibrated from the model score and the decision type, but it does not use content attributes. The attribute-conditioned estimate $\hat{p}^{\text{attr}}_{i,d}$ also conditions on the selected content descriptor. The two-estimator instantiation is

$$\hat{p}^{\text{adapt}}_{i,d} = (1 - w_{i,d})\hat{p}^{\text{score}}_{i,d} + w_{i,d}\hat{p}^{\text{attr}}_{i,d}, \qquad w_{i,d} \in [0, 1]. \tag{3}$$

Figure 1 visualizes this two-estimator case.

Let $u_{i,d}$ denote the decision-time features available to the gate. A parameterized gate produces the mixture weight as

$$w_{i,d} = \sigma(f_\phi(u_{i,d}))\,, \tag{4}$$

where σ bounds the weight to $[0, 1]$. The constant candidates $w_{i,d} = 0$ and $w_{i,d} = 1$ recover the score-only fallback and the always-contextual estimator, respectively. To learn without using held-out evaluation labels, we first generate out-of-fold component predictions $\tilde{p}^{\text{score}}_{i,d}$ and $\tilde{p}^{\text{attr}}_{i,d}$ on calibration data. Parameterized gate candidates minimize calibration-only Brier score,

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{|\mathcal{C}|} \sum_{(i,d)\in\mathcal{C}} \left[(1 - w_\phi(u_{i,d}))\tilde{p}^{\text{score}}_{i,d} + w_\phi(u_{i,d})\tilde{p}^{\text{attr}}_{i,d} - Y_{i,d}\right]^2, \tag{5}$$

where $\mathcal{C}$ contains calibration example–decision pairs. We select among the constant and parameterized candidates by out-of-fold Brier score, then refit the selected component estimators and gate on all calibration rows. Held-out evaluation labels are never used to learn or select the weights.

More generally, the same framework can combine M estimators $\{\hat{p}^{(m)}_{i,d}\}_{m=1}^{M}$ with nonnegative example-dependent weights that sum to one. The two-estimator version is the minimal case needed to test whether content context adds information beyond a decision-level confidence estimate. Fallback is modality-specific: text support weighting forces exact score-only fallback for wholly unseen context, whereas image experiments retain score-only as a candidate but impose no hard per-example fallback. This gives a direct decision-level comparison with non-contextual confidence baselines and a probability-level comparison through Brier score and ECE. Equation (3) is not limited to linear context features: the candidate estimators or gate features can use richer encoders, metadata, embeddings, or additional calibrators when calibration data support them. Section 5 gives the modality-specific gate features and architectures used in our experiments.

The framework is also compatible with traditional supervised calibration targets. In the binary-decision setting used here, the target is the decision label $Y_{i,d}$, so $\hat{p}_{i,d}$ estimates the probability that the decision should be positive. If a task instead provides labels for whether a proposed decision is correct, the same form can calibrate the correctness indicator. If direct ground-truth confidence labels are available, such as human-estimated probabilities or externally measured success rates, the target can be replaced by that confidence value and learned with a proper scoring loss. Conditional confidence calibration therefore extends standard post-hoc calibration rather than replacing it.

5 EXPERIMENTAL SETUP

The experiments test whether context improves confidence calibration relative to matched non-contextual global and decision-specific confidence baselines. The safety datasets provide fine-grained category labels, so every example yields category-level binary decisions. We select calibration policies only on calibration rows, freeze the selected rules, and evaluate once on held-out rows. Official train/test splits are used when available; otherwise we construct a fixed disjoint calibration/evaluation partition before policy selection. The same split is used for all methods within a comparison. The safety evaluation uses SPA-VL (Zhang et al., 2024), MM-SafetyBench++ (Zhang et al., 2026), UnsafeBench (Qu et al., 2024), WildGuardMix (Han et al., 2024), Aegis 2.0 (Ghosh et al., 2025), BeaverTails (Ji et al., 2023), and SALAD-Bench (Li et al., 2024).

Across 10 datasets and 11 evaluation settings, the experiments use 26,513 calibration rows and 27,355 held-out rows. Appendix Table 5 gives modalities, split protocols, and safety/non-safety subtotals. Matched model-size variants for image and image-text data and attribute-source variants for text produce 14 primary comparisons: six image/image-text comparisons and eight text comparisons. Separate model-family evaluations add eight comparisons on two image-text datasets.

Models and scores. All main evaluation rows across the seven safety datasets use the same Gemma 4 E4B checkpoint (`google/gemma-4-E4B-it`). Matched scaling runs use Gemma 4 12B (`google/gemma-4-12B-it`) only on SPA-VL validation and MM-SafetyBench++, where the same prompt, split, attributes, and calibration protocol are available (Gemma Team, 2026). Qwen3-VL and InternVL3.5 evaluations at 4B and 8B use the same held-out rows, attributes, representations, scores, and selection protocol on these datasets (Bai et al., 2025; Wang et al., 2025). Confidence is normalized affirmative versus negative answer-option probability mass under a fixed sequential multi-label prompt with one binary category-presence question per category. The discrete baseline uses generated decisions. Appendix Table 6 gives full scoring details.

Attributes and controls. Image and image-text attributes are extracted by a separate VLM annotation pass with a fixed schema covering visual realism, text load, subject family, scene privacy, human configuration, and action dynamics. Text experiments use rule-based content schemas and a standardized prompt-only LLM attribute extractor. Appendix A.7 reports attribute provenance and support, while Appendix A.8 reports metadata-only, shuffled, and support-matched random controls.

Adaptive gate implementation. Image and image-text gates use frozen SigLIP embeddings with bounded linear or small MLP candidates; text gates use a bounded logistic model over calibrated probabilities, score diagnostics, mapped attributes, and support features. Constant score-only and always-contextual candidates are also considered. All candidates are selected by calibration-only out-of-fold Brier score and refit before held-out evaluation. Appendix A.3 gives the safety estimators, features, support rules, and selection protocol; Appendix A.4.1 gives the non-safety models, features, candidate estimators, cross-fitting, and score extraction protocols.

Policy selection and metrics. Global, per-category, attribute-guided, and adaptive thresholds are selected from calibration data. Attribute-guided overrides require sufficient support and inner-calibration improvement; adaptive thresholds use out-of-fold probabilities. Primary endpoints are Macro BA and Brier score against calibrated decision-level baselines. Micro BA, F1, ECE, exact full-label-set accuracy, Hamming accuracy, and sample-level Jaccard are secondary diagnostics. Aggregate means are unweighted across matched comparisons; relative changes are computed within each comparison before averaging.

Statistical testing. Adaptive comparisons use paired bootstrap resampling for Macro BA, Macro F1, exact full-label-set accuracy, Micro F1, and Brier score with Holm correction[1] Decision-metric families span all 14 primary comparisons; Brier correction is applied separately across the six image/image-text comparisons and eight text comparisons. For each additional model family, Macro BA and Brier tests use separate four-comparison Holm families. ECE, Hamming accuracy, and

[1]Holm's step-down correction controls the family-wise probability of one or more false rejections within each metric/baseline comparison family (Holm, 1979).

sample-level Jaccard are reported descriptively or post hoc. Appendix A gives complete split, implementation, ablation, and statistical details.

6 RESULTS

6.1 ADAPTIVE CALIBRATION IMPROVES HELD-OUT DECISIONS

Table 2 summarizes the 14 primary Gemma 4 model-family comparisons by modality: E4B across all seven safety datasets plus matched 12B runs on SPA-VL validation and MM-SafetyBench++. Relative to the per-category policy, adaptive calibration improves the primary decision metric, Macro BA, by 5.3% on image and image-text tasks and by 5.7% on text tasks. The supporting label-wise F1 metrics move in the same direction: Macro F1 improves by 3.8% on image/image-text evaluations and by 11.0% on text evaluations, while Micro F1 improves by 6.0% and 20.5%. Adaptive calibration also improves Micro BA by 6.6% on image/image-text tasks and 6.7% on text tasks, showing that the decision gain is visible under both category-balanced and pooled views.

Modality	Method	Macro BA	Micro BA	Macro F1	Micro F1	Prob. Brier
Image	Discrete	0.723	0.713	0.414	0.440	–
Image	Global threshold	0.736	0.727	0.413	0.426	
Image	Per-cat thresholds	0.741	0.732	0.424	0.449	0.074
Image	Attr.-guided thresholds	0.742	0.734	0.431	0.458	
Image	Adaptive	**0.781**	**0.781**	**0.440**	**0.476**	**0.069**
Text	Discrete	0.752	0.751	0.409	0.359	–
Text	Global threshold	0.779	0.775	0.366	0.318	
Text	Per-cat thresholds	0.779	0.782	0.386	0.354	0.037
Text	Attr.-guided thresholds	0.779	0.782	0.389	0.357	
Text	Adaptive	**0.823**	**0.834**	**0.429**	**0.427**	**0.035**

Table 2: Mean decision and probability results. Image means cover three datasets (four settings; six comparisons), and text means cover four datasets (eight comparisons). Macro BA is primary; Micro BA and F1 are supporting diagnostics. The three threshold policies share calibrated score-only probabilities, hence the merged Brier cell. Adaptive produces distinct probabilities. Lower Brier is better; bold marks the best.

The paired-bootstrap summary in Appendix Table 14 shows the same pattern: adaptive calibration improves Macro BA and Macro F1 in 12 of 14 primary comparisons and Micro F1 in 13 of 14 primary comparisons, with no Holm-significant regressions for these label-wise metrics. Appendix Table 17 reports the comparison-level Macro BA intervals and adjusted p-values. Appendix Tables 15 and 16 give evaluation-level point estimates and Holm-corrected directions for all decision metrics. Exact full-label-set accuracy is stricter because any single category error makes a row incorrect; it improves in five of 14 primary comparisons and should be optimized separately when exact set recovery is the deployment objective.

6.2 DATASET AND MODEL-SIZE EFFECTS

Table 3 shows the dataset-level Macro BA results behind the averages. Confidence-based policies usually improve over discrete generation, confirming that answer-option probabilities contain useful reliability information. Adaptive calibration is the strongest aggregate policy: it improves over the per-category policy on all image/image-text datasets and on three of four text datasets. SPA-VL official test is the strictest transfer check because policies are selected on validation rows and applied unchanged to the official test set; the gain is positive but smaller than on the validation holdout.

Dataset	Discrete	Global	Per-cat	Attr.-guided	Adaptive	vs disc.	vs per-cat
Image and image-text evaluations							
MM-SafetyBench++	0.738	0.762	0.768	0.769	**0.798**	+8.15%	+3.92%
SPA-VL validation	0.715	0.719	0.730	0.737	**0.756**	+5.75%	+3.60%
SPA-VL official test	0.695	0.687	0.701	0.702	**0.704**	+1.27%	+0.41%
UnsafeBench	0.722	0.746	0.744	0.737	**0.831**	+15.08%	+11.73%
Text-only evaluations							
WildGuardMix	0.748	0.824	0.828	**0.831**	0.827	+10.49%	-0.19%
Aegis 2.0	0.733	0.757	0.734	0.734	**0.822**	+12.04%	+11.94%
BeaverTails	0.831	0.837	0.838	0.840	**0.845**	+1.66%	+0.79%
SALAD	0.704	0.708	0.724	0.722	**0.774**	+9.86%	+6.83%

Table 3: Dataset-level Macro BA. Bold marks the best absolute BA in each row; relative columns compare adaptive calibration with discrete generation and the matched per-category policy.

The same pattern appears across model sizes and domains. Appendix Table 13 compares Gemma 4 E4B with Gemma 4 12B where matched runs are available. Adaptive calibration improves over the per-category policy for both model sizes on SPA-VL validation and MM-SafetyBench++, indicating that the effect is not limited to the smaller model. Beyond safety, it lowers Brier by 3.2–12.1% and improves balanced accuracy by 2.0–2.8% on MMLU-Pro, ScienceQA, and CLINC150; all three Brier gains are Holm-significant (Appendix Tables 8 and 9).

Across Qwen3-VL and InternVL3.5 at 4B and 8B, Macro BA improves by 3.89–7.79% and score-only Brier falls in all eight comparisons, all Holm-significant ($p_{\mathrm{adj}} = 0.00080$); Table 4 summarizes all three families.

Family	Sizes	Comparisons	Macro BA gain	Brier reduction	Holm sig. BA/Brier
Gemma 4	E4B, 12B	4	3.60–8.59%	5.21–11.04%	4/4, 4/4
Qwen3-VL	4B, 8B	4	4.66–7.79%	7.22–8.45%	4/4, 4/4
InternVL3.5	4B, 8B	4	3.89–6.52%	6.52–8.67%	4/4, 4/4
All	–	12	3.60–8.59%	5.21–11.04%	12/12, 12/12

Table 4: Cross-family transfer summary. BA gains are relative to per-category thresholding; Brier reductions compare adaptive with calibrated score-only probabilities. Significance counts use within-family Holm correction. Appendix Tables 13, 11, and 12 report exact values.

6.3 Where Context Helps

Aggregate scores do not fully show how context changes decisions. The attribute-guided policy changes the per-category estimate only when a content group is sufficiently represented in calibration data and improves the calibration-only selection objective. Appendix Table 21 gives descriptive examples from eligible groups to illustrate where the local confidence-error pattern differs from the category average. These examples are not the basis for the aggregate claims; Appendix A.10 defines the support rule without held-out labels and summarizes every eligible group. On the SPA-VL human-centered information-safety group, Macro BA rises by 7.43% relative and F1 by 60.96% relative over the per-category policy, with 89 fewer false positives and no true-positive loss. This supports the central mechanism: context is useful when it identifies a repeatable local calibration error. Appendix A.8 reports controls that replace real content attributes with metadata-only fields, shuffled attributes, or random buckets. Capacity-matched controls favor real image context over shuffled and random features in all six Brier comparisons, with five significant gains after Holm correction for each control. Text gains are smaller and absent against the score-diagnostics-only gate on average (Appendix Table 20).

The probability-calibration evidence uses Brier score as the primary outcome. Brier is a proper scoring rule over continuous decision probabilities, so it rewards both calibrated probabilities and useful separation between easier and harder examples without requiring arbitrary confidence bins.

Against calibrated score-only probabilities, adaptive calibration lowers Brier score in all six image and all eight text primary comparisons. The mean comparison-level relative reduction is 6.3%, with Holm-significant reductions in 6/6 image and 6/8 text comparisons (Appendix Table 10). We keep aggregate 10-bin ECE as a secondary diagnostic in the appendix because it depends on binning and the adaptive estimator is selected for Brier score, not ECE. Comparisons with unchanged raw probabilities are kept separate from calibrated score-only comparisons.

Appendix A.9 shows representative reliability curves for raw, per-category histogram-calibrated, and adaptive probabilities. ECE-targeted attribute calibration also lowers raw per-category ECE in 200 of 204 well-supported image/image-text content groups (Appendix A.10). The calibration takeaway mirrors the decision result: contextual information improves the selected objective, and metric-specific variants should be used when the deployment target is ECE rather than Brier score.

7 Discussion and Limitations

The findings support the central hypothesis that context improves calibration when it explains residual variation in outcome probability after conditioning on model evidence and decision identity. Conditional confidence calibration therefore exposes a support–specificity tradeoff: a contextual estimator can correct systematic local error, but becomes less reliable as the relevant calibration subset becomes smaller. Adaptive calibration operationalizes this tradeoff by allowing a contextual estimate to modify, rather than replace, a better-supported decision-level estimate.

The controls suggest that the observed gains are not explained solely by additional estimator capacity: meaningful image context is more effective than shuffled or random features. The weaker and more metric-dependent text results also indicate that context is not intrinsically useful. Its value depends on whether the representation captures stable reliability differences, whether those differences have sufficient calibration support, and whether they align with the selected objective. The score-only candidate and support-aware gate provide a principled fallback when these conditions are not met.

The cross-family and non-safety evaluations broaden the evidence beyond the primary Gemma safety setting, but they do not imply universal effect sizes. Decision quality, Brier score, ECE, and exact-set accuracy measure distinct properties and need not improve together; the calibration objective should therefore be chosen for the intended use. More generally, the results support conditional calibration as a design framework rather than a claim that every available context variable should be used.

Limitations: The evaluation centers on fine-grained safety decisions. Three non-safety tasks and three model families broaden coverage, but the additional-family tests cover only two image-text datasets. Broader evaluation across domains and deployment conditions is needed to establish how reliably the learned conditioning transfers. The method requires decision-time context that reflects supported reliability differences, and can be affected by extraction error, schema choice, or distribution shift. Coarser schemas, hierarchical pooling, and automatic fallback may help sparse groups. Because post-hoc calibration does not change the underlying model, deployment still requires monitoring and an objective matched to the application.

8 Conclusion

We propose conditional confidence calibration, a post-hoc framework that estimates confidence from model evidence, decision identity, and context available when the decision is made. This matters because global or decision-level calibration cannot correct repeatable reliability differences within the same decision category. Adaptive calibration learns when a contextual estimate should modify a better-supported fallback without updating the base model. Across seven safety datasets, it improves Macro BA in 12/14 primary comparisons and Brier in 14/14, with a 6.3% mean reduction. Cross-family, supported-group, and non-safety results show transfer. The practical rule is to use context for supported reliability differences and otherwise retain decision-level confidence; exact-set accuracy and ECE remain separate objectives.

AI USE STATEMENT

The authors used AI-assisted tools for drafting support, code assistance, LaTeX editing, experiment monitoring, and consistency checks. The authors reviewed and verified the manuscript content, experimental claims, and final submission materials, and remain responsible for the work.

REPRODUCIBILITY STATEMENT

The main text defines the scoring, calibration, policy-selection, and held-out evaluation protocol. Appendix A records split provenance, attribute extraction, support counts, leakage controls, statistical testing, and artifact provenance for the reported results. Together, these sections document confidence-score construction, global and decision-level calibration, attribute-conditioned calibration, adaptive gating, metric selection, and bootstrap evaluation. Datasets, model outputs, and extracted attribute annotations are not included with the manuscript.

ETHICS STATEMENT

The experiments use safety datasets that contain harmful or sensitive content. They are used only for evaluation and calibration analysis, not to train models to produce harmful outputs. Attribute extraction is restricted to context available when the safety decision is made and excludes labels, correctness indicators, confidence scores, and evaluation metrics. For safeguards that classify an already-generated candidate response, that response is part of the available decision-time context.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 295–302, 2020. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. Rethinking LLM uncertainty: A multi-agent approach to estimating black-box model uncertainty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12349–12375, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.660. URL https://aclanthology.org/2025.findings-emnlp.660/.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017. URL https://arxiv.org/abs/1705.08500.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192*, 2019. URL https://arxiv.org/abs/1901.09192.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5992–6026, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.306. URL https://aclanthology.org/2025.naacl-long.306/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-0261. URL https://mlanthology.org/neurips/2024/han2024neurips-wildguard/.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. URL https://arxiv.org/abs/1610.02136.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. *arXiv preprint arXiv:2307.04657*, 2023. URL https://arxiv.org/abs/2307.04657.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *arXiv preprint arXiv:2012.00955*, 2020. URL https://arxiv.org/abs/2012.00955.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14919–14926, 2023. doi: 10.1609/aaai.v37i12.26742. URL https://ojs.aaai.org/index.php/AAAI/article/view/26742.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. *arXiv preprint arXiv:1805.12317*, 2018. URL https://arxiv.org/abs/1805.12317.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1311–1316. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131/.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024. URL https://arxiv.org/abs/2402.05044.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022. URL https://arxiv.org/abs/2205.14334.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430. URL https://mlanthology.org/icml/2005/niculescumizil2005icml-predicting/.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–41, 2019. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999. URL https://cir.nii.ac.jp/crid/1573387450082775168?lang=en.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. UnsafeBench: Benchmarking image safety classifiers on real-world and AI-generated images. *arXiv preprint arXiv:2405.03486*, 2024. URL https://arxiv.org/abs/2405.03486.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. URL https://www.jmlr.org/papers/v9/shafer08a.html.

Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *Computer Vision – ECCV 2022*, volume 13673 of *Lecture Notes in Computer Science*, pp. 555–569. Springer, 2022. doi: 10.1007/978-3-031-19778-9_32. URL https://doi.org/10.1007/978-3-031-19778-9_32.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.01574. Datasets and Benchmarks Track.

Duygu Nur Yaldiz, Evangelia Spiliopoulou, Zheng Qi, Siddharth Varia, Srikanth Doss, and Nikolaos Pappas. Balancing classification and calibration performance in decision-making LLMs via calibration aware reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12537–12553, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.610. URL https://aclanthology.org/2026.findings-acl.610/.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151. URL https://dl.acm.org/doi/10.1145/775047.775151.

Ce Zhang, Jinxi He, Junyi He, Katia Sycara, and Yaqi Xie. Evolving contextual safety in multi-modal large language models via inference-time self-reflective memory. *arXiv preprint arXiv:2603.15800*, 2026. URL https://arxiv.org/abs/2603.15800.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. SPA-VL: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*, 2024. URL https://arxiv.org/abs/2406.12030.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022. URL https://arxiv.org/abs/2210.00045.

A Appendix

This appendix keeps the reviewer-facing details needed to interpret the main claims: split construction, scoring, adaptive comparison-level statistics, attribute provenance, ablation controls, supported-group calibration error, and three non-safety evaluations.

A.1 Extended Related Work

Classical calibration converts model scores into probabilities that match empirical correctness. Platt scaling, isotonic regression, Bayesian binning, and temperature scaling are standard post-hoc examples (Platt, 1999; Zadrozny & Elkan, 2002; Pakdaman Naeini et al., 2015; Guo et al., 2017). These methods established that classification accuracy and probability quality are distinct objectives (Niculescu-Mizil & Caruana, 2005). More recent methods increase expressive power by learning prediction- or sample-specific temperatures rather than one global temperature. Parameterized Temperature Scaling uses a neural calibrator to produce prediction-specific temperatures, while Sample-Dependent Adaptive Temperature Scaling learns a separate temperature for each input (Tomani et al., 2022; Joy et al., 2023). These methods are close in spirit because they make post-hoc calibration input dependent. In our notation, classical calibrators learn $g(S)$ or $g(S, d)$, whereas these adaptive temperature methods enrich the mapping from model evidence S.

Our framework instead makes the conditioning variables explicit and includes decision identity and semantically interpretable pre-decision context. The proposed adaptive estimator also retains a better-supported decision-specific fallback when contextual estimates are sparse. We keep the post-hoc advantage of leaving the base model fixed and test this support-aware conditioning on prompted foundation-model decisions rather than only classifier-logit scaling.

Neural uncertainty estimation uses signals such as maximum softmax probability, margins, entropy, Bayesian approximations, or ensembles (Hendrycks & Gimpel, 2016; Gal & Ghahramani, 2016;

Lakshminarayanan et al., 2017). Modern neural networks can remain overconfident despite high accuracy (Guo et al., 2017), and calibration can shift under domain change (Desai & Durrett, 2020). These works provide useful model evidence S. Our question is different: given such evidence, what additional conditioning information should a post-hoc confidence estimator use before it is frozen and evaluated on held-out decisions?

Conditioning calibration on identifiable groups has clear precedents and is the closest line of work to our framework. Multicalibration requires calibrated predictions over many intersecting subpopulations, while multiaccuracy corrects systematic group-level prediction errors (Hébert-Johnson et al., 2018; Kim et al., 2018). Multicalibration is therefore the closest probability-calibration antecedent. Mondrian or subgroup-aware conformal prediction provides a related coverage-based view by partitioning calibration examples using known categories or attributes (Shafer & Vovk, 2008). These methods motivate our support requirement: small groups need enough calibration examples before local estimates are trustworthy.

Accordingly, our novelty claim is not that conditional or group-aware calibration is unprecedented. Our contribution is to formulate the conditioning choice for prompted foundation-model decisions, use pre-decision content descriptors rather than protected-group membership or a fixed partition, and introduce an adaptive estimator that combines a local contextual estimate with a better-supported decision-specific estimate instead of committing to one granularity. We then test this support–specificity tradeoff across held-out LLM and VLM decisions.

LLM uncertainty work shows that token probabilities, generated self-evaluations, verbalized confidence, and black-box multi-agent estimates can all provide imperfect but useful reliability signals (Jiang et al., 2020; Kadavath et al., 2022; Lin et al., 2022; Feng et al., 2025). Calibration-aware reinforcement learning offers a training-time route to balancing decision accuracy and calibration (Yaldiz et al., 2026), while sequence-likelihood calibration targets open-ended generation quality (Zhao et al., 2022). Our setting is different: we do not retrain the model, and we do not score open-ended generation quality directly. We calibrate prompted decisions using option probability mass and evaluate whether adding context improves the frozen post-hoc estimator.

Confidence estimates can also support selective prediction and routing (Geifman & El-Yaniv, 2017; 2019). This paper isolates the calibration layer before adding abstention or human review. The resulting question is narrower and experimentally cleaner: does conditioning confidence on decision type and content context improve the decision that would otherwise be made?

A.2 Dataset Splits and Scoring

Table 5 summarizes the datasets, modalities, and split sizes. Policies are selected on calibration rows only and then applied unchanged to held-out rows. ScienceQA and CLINC150 use official validation/test splits; MMLU-Pro uses a frozen held-out partition.

A.3 Safety-Method Reproducibility

A.4 Non-Safety Evaluations

We evaluate conditional calibration beyond safety classification on three non-safety tasks. ScienceQA (Lu et al., 2022) and CLINC150 (Larson et al., 2019) use their official validation splits for calibration and model selection and their official test splits for the reported results. MMLU-Pro (Wang et al., 2024) uses a frozen held-out partition because its public release does not provide the same validation–test protocol. No reported evaluation rows are used to fit a calibrator, select a threshold, or choose a model. For ScienceQA and MMLU-Pro, the calibrated outcome is whether the model's multiple-choice answer is correct. For CLINC150, it is whether the predicted intent is correct. The strongest score-only calibrator selected on calibration data is the baseline for each dataset.

Adaptive calibration lowers held-out Brier score on all three datasets by 3.2–12.1% relative to the strongest calibrated score-only baseline. It also improves balanced accuracy by 2.0–2.8% and improves NLL, AURC, and AUROC in every evaluation. ECE improves on ScienceQA and CLINC150 but increases from .0133 to .0157 on MMLU-Pro, illustrating the sensitivity of this binned diagnostic even when the proper scoring rules and ranking metrics improve.

Dataset/evaluation	Modality	Calibration rows	Held-out rows
Safety tasks			
SPA-VL validation	image-text	3,000	2,000
SPA-VL official test	image-text	3,000	530
MM-SafetyBench++	image-text	1,560	1,064
UnsafeBench	image	500	500
WildGuardMix	text	600	500
Aegis 2.0	text	900	600
BeaverTails	text	3,000	2,000
SALAD	text	3,000	2,000
Non-safety tasks			
MMLU-Pro	text	3,612	8,420
ScienceQA	image-text	4,241	4,241
CLINC150	text	3,100	5,500
Safety subtotal		15,560	9,194
Non-safety subtotal		10,953	18,161
All evaluations		26,513	27,355

Table 5: Datasets, modalities, and evaluation sizes. The study contains 10 unique datasets and 11 evaluation settings because SPA-VL validation and official test are evaluated separately. Calibration and held-out rows are disjoint within every setting.

Role	Model ID	Backend / precision	Score source
Main runs	`google/gemma-4-E4B-it`	vLLM, BF16	selected answer-option log probabilities
Scaling runs	`google/gemma-4-12B-it`	vLLM, BF16	selected answer-option log probabilities

Table 6: Model and scoring configuration. Both models are from the Gemma 4 family (Gemma Team, 2026). Each category-level decision is scored by summing affirmative answer-option probability mass and normalizing by affirmative plus negative probability mass at the corresponding answer position.

Component	Image and image-text	Text
Component estimators	Per category, score-only candidates are raw, Platt, beta, isotonic, and 5/10/15/20-bin histogram calibration. Contextual candidates are additive and score-interaction logistic models. Logistic penalties are $10^{-2}, 10^{-1}, 1, 10, 10^{2}$.	The same score-only and contextual candidate families, histogram bins, and penalty grid are selected independently for each category and attribute source.
Gate and features	$p_s + w(z)(p_a - p_s)$, where z is a frozen SigLIP So400m Patch14-384 embedding (1152 dimensions, L2 normalized, then a fixed Gaussian projection to 64 dimensions). Candidates are exact score-only, always-contextual, and linear or 16-hidden-unit MLP gates with $w_{\max} \in \{.5, 1\}$, L2 $= .001$, and usage penalty $= .0005$.	$p_s + w(x)(p_a - p_s)$, with score, contextual estimate, their signed and absolute difference, entropy, distance from .5, one-hot attributes, median support, and unseen-value fraction. Candidates use $w_{\max} \in \{0, .25, .5, .75, 1\}$; the main grid uses L2 $= 1$ and usage penalty $= .0005$, with additional $(w_{\max}, \text{L2})$ settings $(.5, .1)$ and $(1, .1)$ without a usage penalty.
Cross-fitting	Five folds, deterministically stratified by the binary label within each category, produce out-of-fold component and gate probabilities. Selection minimizes category-level out-of-fold Brier score.	Five label-stratified folds select component estimators. Gate selection uses four grouped folds over the component-fold identifiers, preventing a gate row from using in-fold component predictions. Selection minimizes group-macro out-of-fold Brier score.
Support and fallback	Attribute values require at least 10 calibration occurrences to enter the contextual estimator. The exact score-only candidate is always present; unsupported values receive no attribute indicator. The visual gate is bounded but has no additional per-example hard support cutoff.	Attribute values require 10 calibration occurrences in the contextual estimator and 20 for full gate support. Gate weight is multiplied by $(1 - u)\min(1, m/20)$, where u is the unseen-field fraction and m is median value support. Thus wholly unseen context enforces exact score-only fallback; a zero-weight candidate is also required.
Attribute extractor	`google/gemma-4-E4B-it` produces deterministic JSON attributes from the image, user text, and a fixed label-free visual schema. The adaptive gate itself uses frozen `google/siglip-so400m-patch14-384`.	`google/gemma-4-E4B-it` produces deterministic JSON attributes from prompt text and a fixed label-free schema. BeaverTails' rule-based variant also describes the candidate response, which is an input to the safeguard and is available before its safety decision; its LLM-extracted variant is prompt-only.

Table 7: Reproducibility details for the safety experiments. Here p_s and p_a denote the selected score-only and contextual probability estimators. Exact prompts, checkpoint revisions, seeds, source hashes, and per-category selections are retained in the supplementary configurations and extraction reports.

Table 8: Held-out non-safety results. BA and F1 measure prediction of the binary correctness outcome, rather than native task accuracy. Lower is better for Brier, NLL, AURC, and ECE; higher is better for BA, F1, and AUROC. The relative-change rows compare adaptive calibration with the score-only baseline. ECE is included as a binned diagnostic and was not the selection objective.

Dataset	Method	BA↑	F1↑	Brier↓	NLL↓	AURC↓	AUROC↑	ECE↓
MMLU-Pro	Score-only	.715	.641	.1788	.5380	.3598	.7799	**.0133**
	Adaptive	**.730**	**.677**	**.1730**	**.5225**	**.3501**	**.7945**	.0157
	Relative change	+2.05%	+5.64%	−3.23%	−2.89%	−2.68%	+1.86%	+18.22%
ScienceQA	Score-only	.859	.927	.0824	.2482	.0249	.9082	.0168
	Adaptive	**.883**	**.934**	**.0729**	**.2204**	**.0212**	**.9317**	**.0139**
	Relative change	+2.83%	+0.78%	−11.50%	−11.19%	−14.77%	+2.58%	−17.30%
CLINC150	Score-only	.832	.819	.1196	.3738	.2508	.9141	.0472
	Adaptive	**.852**	**.842**	**.1051**	**.3369**	**.2327**	**.9304**	**.0234**
	Relative change	+2.34%	+2.76%	−12.13%	−9.88%	−7.23%	+1.79%	−50.37%

Table 9: Paired bootstrap evidence for the adaptive Brier-score improvement over the strongest calibrated score-only baseline. Confidence intervals are 95% percentile intervals from 10,000 paired resamples; p-values use Holm correction across the three non-safety datasets.

Dataset	Absolute reduction	95% CI	Holm p
MMLU-Pro	.00577	[.00418, .00732]	.0032
ScienceQA	.00947	[.00699, .01201]	.0012
CLINC150	.01451	[.01178, .01727]	.0012

The native task accuracies were 85.62% for ScienceQA and 45.20% for the two-stage CLINC150 intent classifier. These values characterize the fixed base model and are not changed by post-hoc calibration. The MMLU-Pro base answer accuracy was 40.49%. ScienceQA used 4,241 validation and 4,241 test examples; CLINC150 used 3,100 validation and 5,500 test examples; MMLU-Pro used 3,612 calibration and 8,420 held-out examples.

A.4.1 Reproducibility Details

Models and score extraction. All three evaluations use `google/gemma-4-E4B-it` in BF16 through vLLM. ScienceQA uses `derek-thomas/ScienceQA`; CLINC150 uses the official `clinc/oos-eval` mapping and the `clinc/clinc_oos` dataset mirror; MMLU-Pro uses `TIGER-Lab/MMLU-Pro`.

For ScienceQA and MMLU-Pro, the model emits one multiple-choice answer symbol. We aggregate all validated single-token variants of each valid symbol at the first generated position and normalize their probability mass over the listed options. The resulting vector provides the selected-option probability, normalized entropy, top-two margin, option count, and retained answer-symbol mass. The calibrated outcome is whether the selected answer is correct. CLINC150 uses analogous symbol-probability extraction at both stages of the hierarchy described below. No gold answer, correctness indicator, or evaluation statistic is provided to the model or attribute extractor.

Context variables. MMLU-Pro uses the dataset subject and six attributes extracted from the question and options: reasoning type, answer format, quantitative requirement, terminology level, reasoning depth, and question context. ScienceQA uses inference-time dataset context (modality, subject, grade band, topic, and category) together with model-extracted visual format, reasoning type, quantitative requirement, context dependence, and terminology level. CLINC150 uses a deterministic utterance-length bucket and model-extracted utterance form, lexical specificity, named-entity presence, numeric/date/time reference, context dependence, ambiguity, and urgency. Extraction uses deterministic JSON generation by the evaluated Gemma checkpoint and sees only inference-time content: the question, options, hint, and image for ScienceQA; the question and options for MMLU-Pro; and the utterance for CLINC150. Invalid fields map to declared fallback values and are counted as extraction failures.

Candidate calibrators and selection. For every dataset, we compare raw confidence, global isotonic calibration, and a supported group-specific isotonic calibrator (subject for ScienceQA and MMLU-Pro; domain for CLINC150). The learned score-only and contextual candidate sets use regularized logistic regression with candidate values $\{0.03, 0.1, 0.3, 1\}$, histogram gradient boosting with candidate values $\{1, 5, 20\}$, and regularized multilayer perceptrons with candidate values $\{0.01, 0.1\}$. Candidate selection minimizes out-of-fold Brier score, breaking ties by negative log-likelihood and area under the risk–coverage curve. Histogram gradient boosting with value 20 is the selected score-only estimator for ScienceQA and CLINC150. For MMLU-Pro, global isotonic calibration is the strongest calibrated score-only baseline, and logistic regression with value 0.1 is the selected contextual estimator.

Adaptive estimator and cross-fitting. Each adaptive estimator is a regularized softmax mixture of three probability experts. ScienceQA combines global isotonic, subject-specific isotonic, and score-plus-context estimates; CLINC150 combines global isotonic, domain-specific isotonic, and score-plus-context estimates; MMLU-Pro combines global isotonic, subject-specific isotonic, and score-plus-context estimates. The gate always receives an intercept, log group support, and disagreement

among its experts. The context-aware gate additionally receives score diagnostics (selected score, entropy, and margin, as available) and one-hot context variables. MMLU-Pro also supplies support for the observed attribute combination. We compare constant, support-aware, and context-aware gates with regularization selected on calibration data. The context-aware gate with regularization 0.1 is selected for all three datasets.

All estimator, hyperparameter, gate, and decision-threshold choices use five-fold out-of-fold predictions from calibration data only. ScienceQA folds are stratified by correctness and broad subject/modality; CLINC150 folds are stratified by correctness and intent/domain; MMLU-Pro folds are stratified by subject and correctness. After selection, each component is refit on all calibration rows and applied once, unchanged, to held-out rows. ScienceQA and CLINC150 use their complete official validation splits for selection and their official test splits for evaluation. MMLU-Pro uses a deterministic 30/70 split of the public test set, stratified by subject and gold answer option with seed `20260923`; its result is therefore a frozen held-out evaluation rather than an official-test transfer result.

CLINC150 hierarchy. CLINC150 avoids treating 151 intent names as atomic answer tokens. Stage 1 selects one of ten official domains or out-of-scope using an 11-way symbol prompt. If an in-scope domain is selected, Stage 2 chooses among that domain's 15 intents using a second symbol prompt. For an in-scope prediction, final confidence is

$$p(\widehat{d} \mid x)\, p(\widehat{y} \mid \widehat{d}, x),$$

where $\widehat{d}$ is the predicted domain and $\widehat{y}$ is the predicted intent. For an out-of-scope prediction, final confidence is $p(\mathrm{OOS} \mid x)$. The calibrated binary outcome is exact correctness of the final intent or out-of-scope decision. Domain accuracy is 49.71% and final intent/out-of-scope accuracy is 45.20% on the official test split; the calibration methods do not change these base predictions.

A.5 Adaptive Probability-Calibration Statistical Details

Table 12 reports every matched adaptive-versus-calibrated-score-only probability comparison. These are the comparisons underlying the 6.3% mean comparison-level Brier reduction in the main text; they are distinct from comparisons against unchanged raw probabilities.

Modality	Comparisons	Score-only Brier	Adaptive Brier	Mean rel. reduction	Holm sig.
Image/image-text	6	0.0737	0.0693	6.47%	6/6
Text	8	0.0370	0.0345	6.19%	6/8
All	14	0.0527	0.0494	6.31%	12/14

Table 10: Adaptive probability calibration against the matched calibrated score-only baseline. Mean relative reduction is computed within each comparison before averaging.

Evaluation across vision–language model families. We repeat the frozen evaluation protocol with Qwen3-VL and InternVL3.5 at 4B and 8B parameters (Bai et al., 2025; Wang et al., 2025). Table 11 reports comparisons on the same held-out SPA-VL validation and MM-SafetyBench++ rows used for the primary model-size evaluations.

All eight adaptive-minus-per-category Macro BA confidence intervals exclude zero. All eight adaptive-minus-score-only Brier confidence intervals also exclude zero, and the comparisons remain significant after Holm correction within each model family's four-comparison metric family ($p_{\mathrm{adj}} = 0.00080$). A clean Qwen3-VL calibration replay reproduced the held-out adaptive probabilities and result summaries byte-for-byte. For InternVL3.5 8B, the requested Yes/No log probabilities are complete, but the model often samples its official end token at the answer position. We therefore use the predeclared score-based fallback and do not treat discrete generation from that checkpoint as primary scaling evidence.

Family	Size	Dataset	Per-cat. BA	Adaptive BA	BA gain	Score-only Brier	Adaptive Brier
Qwen3-VL	4B	SPA-VL	0.667	**0.719**	+7.79%	0.1006	**0.0921**
Qwen3-VL	4B	MM-SafetyBench++	0.752	**0.802**	+6.64%	0.0706	**0.0650**
Qwen3-VL	8B	SPA-VL	0.702	**0.735**	+4.66%	0.0942	**0.0874**
Qwen3-VL	8B	MM-SafetyBench++	0.765	**0.814**	+6.36%	0.0671	**0.0622**
InternVL3.5	4B	SPA-VL	0.703	**0.749**	+6.52%	0.0937	**0.0871**
InternVL3.5	4B	MM-SafetyBench++	0.744	**0.791**	+6.27%	0.0721	**0.0674**
InternVL3.5	8B	SPA-VL	0.697	**0.739**	+6.01%	0.0937	**0.0871**
InternVL3.5	8B	MM-SafetyBench++	0.780	**0.811**	+3.89%	0.0692	**0.0632**

Table 11: Evaluation across additional vision–language model families on held-out data. BA gain is the relative improvement of adaptive calibration over per-category thresholding. Brier scores compare adaptive probabilities with the matched calibrated score-only estimator; lower is better. Bold marks the better result in each comparison.

Evaluation	Source	Score-only	Adaptive	Rel. reduction	Paired 95% CI	Holm p
12B MM-SafetyBench++	visual	0.05941	0.05285	11.03%	[-0.00837, -0.00474]	0.0012
12B SPA-VL validation	visual	0.08895	0.08432	5.21%	[-0.00560, -0.00369]	0.0012
E4B MM-SafetyBench++	visual	0.06644	0.06152	7.41%	[-0.00670, -0.00308]	0.0012
E4B SPA-VL official	visual	0.10321	0.10095	2.19%	[-0.00417, -0.00038]	0.0180
E4B SPA-VL validation	visual	0.09033	0.08511	5.77%	[-0.00620, -0.00420]	0.0012
E4B UnsafeBench	visual	0.03367	0.03124	7.21%	[-0.00365, -0.00119]	0.0012
Aegis 2.0	LLM attributes	0.02766	0.02618	5.34%	[-0.00208, -0.00089]	0.0016
Aegis 2.0	rule attributes	0.02783	0.02672	3.99%	[-0.00171, -0.00052]	0.0016
BeaverTails	LLM attributes	0.04521	0.04016	11.16%	[-0.00581, -0.00431]	0.0016
BeaverTails	rule attributes	0.04235	0.04067	3.95%	[-0.00213, -0.00122]	0.0016
SALAD	LLM attributes	0.04465	0.03837	14.07%	[-0.00695, -0.00563]	0.0016
SALAD	rule attributes	0.04466	0.04149	7.11%	[-0.00366, -0.00271]	0.0016
WildGuardMix	LLM attributes	0.03172	0.03076	3.03%	[-0.00188, -0.00007]	0.0736
WildGuardMix	rule attributes	0.03163	0.03135	0.89%	[-0.00090, +0.00033]	0.3560

Table 12: Comparison-level calibrated probability results. The interval is for adaptive minus score-only Brier, so negative values favor adaptive calibration. Holm correction is applied separately within the six image and eight text comparison families.

A.6 Adaptive Thresholded Operational Details

Table 14 summarizes direction and paired-bootstrap significance relative to the per-category confidence policy. The table is placed here because it supports the main decision result but is diagnostic rather than the primary aggregate comparison.

Dataset	Size	Disc. BA	Per-cat BA	Adapt. BA	Adapt. vs disc.	Adapt. vs per-cat
SPA-VL validation	4B	0.715	0.730	**0.756**	+5.75%	+3.60%
SPA-VL validation	12B	0.657	0.684	**0.742**	+12.92%	+8.59%
SPA-VL official test	4B	0.695	0.701	**0.704**	+1.27%	+0.41%
MM-SafetyBench++	4B	0.738	0.768	**0.798**	+8.15%	+3.92%
MM-SafetyBench++	12B	0.813	0.823	**0.854**	+4.93%	+3.72%
UnsafeBench	4B	0.722	0.744	**0.831**	+15.08%	+11.73%

Table 13: Model-size comparison for image and image-text Macro BA. Bold marks the best absolute BA in each row; relative columns compare adaptive calibration with discrete generation and the per-category confidence policy.

Metric	Win rate vs per-cat	Holm-significant gains	Holm-significant losses
Macro BA	86% (12/14)	10/14	0/14
Macro F1	86% (12/14)	7/14	0/14
Exact full-label-set	36% (5/14)	3/14	6/14
Micro F1	93% (13/14)	9/14	0/14

Table 14: Direction and paired-bootstrap significance relative to the per-category confidence policy. Macro BA is the primary decision metric; exact full-label-set accuracy is a stricter diagnostic reported for transparency.

Table 15 gives the matched per-category-policy/adaptive point estimates for all four decision metrics. Exact is strict full-label-set accuracy, not string equality: all thresholded category decisions for a row must match the reference vector.

Evaluation	Source	Macro BA P/A	Macro F1 P/A	Exact P/A	Micro F1 P/A
E4B SPA-VL official	visual	.7007/.7036	.3928/.3941	.1075/.1208	.4186/.4355
E4B MM-SafetyBench++	visual	.7679/.7981	.4820/.4671	.4182/.3816	.4586/.4661
E4B UnsafeBench	visual	.7436/.8308	.3877/.4369	.4780/.4040	.4024/.4351
E4B SPA-VL validation	visual	.7298/.7561	.3598/.4075	.1215/.1630	.4307/.5020
12B SPA-VL validation	visual	.6835/.7423	.3866/.4083	.2675/.2150	.4721/.5024
12B MM-SafetyBench++	visual	.8230/.8536	.5360/.5273	.4314/.3957	.5129/.5167
WildGuardMix	rule	.8284/.8269	.4106/.4379	.2360/.2660	.3536/.3974
WildGuardMix	LLM	.8284/.8257	.4106/.4481	.2360/.0940	.3536/.3466
Aegis 2.0	rule	.7341/.8218	.3827/.4051	.1917/.0967	.3109/.4251
Aegis 2.0	LLM	.7341/.8233	.3827/.4185	.1917/.1517	.3109/.4370
BeaverTails	rule	.8381/.8447	.4607/.4726	.1665/.2085	.4557/.5085
BeaverTails	LLM	.8207/.8404	.4538/.4724	.1465/.2245	.4268/.5212
SALAD	rule	.7243/.7738	.2947/.3624	.0745/.0550	.3106/.3677
SALAD	LLM	.7243/.8309	.2947/.4134	.0745/.0630	.3106/.4108

Table 15: Evaluation-level decision metrics. P/A denotes per-category and adaptive thresholding on the same held-out rows. Higher is better.

Evaluation	Macro BA	Macro F1	Exact set	Micro F1
12B MM-SafetyBench++	gain	n.s.	loss	n.s.
12B SPA-VL validation	gain	gain	loss	gain
E4B MM-SafetyBench++	gain	n.s.	n.s.	n.s.
E4B SPA-VL official	n.s.	n.s.	n.s.	n.s.
E4B SPA-VL validation	gain	gain	gain	gain
E4B UnsafeBench	gain	n.s.	loss	n.s.
Aegis 2.0, LLM	gain	n.s.	n.s.	gain
Aegis 2.0, rule	gain	n.s.	loss	gain
BeaverTails, LLM	gain	gain	gain	gain
BeaverTails, rule	n.s.	n.s.	gain	gain
SALAD, LLM	gain	gain	n.s.	gain
SALAD, rule	gain	gain	loss	gain
WildGuardMix, LLM	n.s.	gain	loss	n.s.
WildGuardMix, rule	n.s.	gain	n.s.	gain

Table 16: Holm-corrected paired-bootstrap direction for adaptive versus per-category thresholding. Gain/loss indicates a significant adaptive increase/decrease at $\alpha = 0.05$; n.s. is not significant.

The strict exact-set result is sensitive to any single category error. As a post-hoc interpretation, adaptive improves Hamming accuracy and sample-level Jaccard in nine of 14 comparisons relative

to per-category thresholding, even though exact-set accuracy improves in only five. These softer diagnostics do not replace the pre-specified exact-set endpoint.

Evaluation	Macro BA delta	Paired 95% CI	Holm p	Significant
E4B SPA-VL validation	+0.0263	[+0.0143, +0.0382]	0.0028	yes
E4B SPA-VL official	+0.0029	[-0.0151, +0.0202]	1.0000	no
E4B MM-SafetyBench++	+0.0302	[+0.0123, +0.0483]	0.0060	yes
E4B UnsafeBench	+0.0874	[+0.0620, +0.1118]	0.0028	yes
12B SPA-VL validation	+0.0587	[+0.0428, +0.0746]	0.0028	yes
12B MM-SafetyBench++	+0.0307	[+0.0177, +0.0443]	0.0028	yes
WildGuardMix, rule	-0.0016	[-0.0128, +0.0088]	1.0000	no
WildGuardMix, LLM	-0.0028	[-0.0198, +0.0140]	1.0000	no
Aegis 2.0, rule	+0.0867	[+0.0615, +0.1087]	0.0028	yes
Aegis 2.0, LLM	+0.0884	[+0.0641, +0.1133]	0.0028	yes
BeaverTails, rule	+0.0066	[-0.0008, +0.0142]	0.3112	no
BeaverTails, LLM	+0.0198	[+0.0100, +0.0302]	0.0028	yes
SALAD, rule	+0.0495	[+0.0389, +0.0600]	0.0028	yes
SALAD, LLM	+0.1066	[+0.0944, +0.1186]	0.0028	yes

Table 17: Adaptive minus per-category Macro BA. Intervals are paired row-bootstrap intervals; Holm correction is across all 14 primary comparisons in this metric/baseline family.

A.7 Attribute Extraction and Coverage

Attributes are pre-classification context signals. Attribute annotators do not receive true labels, generated labels, confidence scores, calibration parameters, correctness indicators, or metric outputs. Table 18 summarizes the extraction mode and the broad support profile.

Dataset	Extraction mode	Inputs visible to annotator	Attribute fields and support notes
SPA-VL	VLM visual annotation plus derived fields	image, user question, fixed visual schema	text load, human configuration, scene privacy, action dynamics, subject family; broad support across several meaningful visual groups
MM-SafetyBench++	VLM visual annotation plus derived fields	image, user question, fixed visual schema	text load, human configuration, scene privacy, action dynamics, subject family; strong public/human skew with fewer accepted local rules
UnsafeBench	VLM visual annotation plus optional neutral metadata	image, optional neutral manifest text, fixed visual schema	text load, human configuration, scene privacy, action dynamics, subject family, visual realism; broad visual support but metric-dependent effects
WildGuardMix	rule-based content annotation; prompt-only LLM annotation	prompt text, fixed text schema	domain, request type, specificity, target type, requested artifact, framing; 22 LLM-parse failures among 1100 rows
Aegis 2.0	rule-based content annotation; prompt-only LLM annotation	prompt text, fixed text schema	intent, domain, actionability, stance, output form, social scope, realism; 16 LLM-parse failures among 1500 rows
BeaverTails	rule-based content annotation; prompt-only LLM annotation	prompt text for LLM check; prompt-response for sharp rule-based variant	intent, domain, actionability, stance, output form, social scope, realism; 40 LLM-parse failures among 5000 rows
SALAD	rule-based content annotation; prompt-only LLM annotation	question text, fixed text schema	intent, domain, actionability, stance, output form, social scope, realism; 46 LLM-parse failures among 5000 rows

Table 18: Attribute provenance and coverage summary. Attributes are available to the safeguard when it makes the final safety decision; for BeaverTails, this may include an already-generated candidate response.

A.8 Attribute Ablation Controls

Table 19 checks whether content-conditioned gains come from meaningful attributes rather than additional local-rule degrees of freedom. Real content attributes describe visible or semantic input properties. Metadata-only fields come from dataset/source bookkeeping, shuffled content attributes preserve marginal counts while breaking example-level alignment, and random buckets preserve support structure without content semantics.

Dataset	Attribute source	Rules	BA rel. change	Macro F1 rel. change
SPA-VL	real content	12	+0.98%	+4.27%
SPA-VL	metadata-only	0	0.00%	0.00%
SPA-VL	shuffled content	0	0.00%	0.00%
SPA-VL	random buckets	26	+0.08%	+0.14%
MM-SafetyBench++	real content	3	+0.10%	+0.61%
MM-SafetyBench++	metadata-only	1	-0.95%	-1.90%
MM-SafetyBench++	shuffled content	0	0.00%	0.00%
MM-SafetyBench++	random buckets	0	0.00%	0.00%
UnsafeBench	real visual content	44	-0.83%	+5.76%
UnsafeBench	metadata-only	17	-1.11%	+2.02%
UnsafeBench	shuffled content	48	-2.07%	+5.44%
UnsafeBench	random buckets	39	-1.83%	+3.14%

Table 19: Attribute ablations relative to the per-category confidence policy for decision-metric policies. Real content attributes produce the clearest gains on SPA-VL and smaller positive effects on MM-SafetyBench++; results are mixed on UnsafeBench, where non-semantic controls also improve Macro F1.

Capacity-matched context controls. We test whether the adaptive method benefits from meaningful context rather than merely from additional model capacity. We compare it with the same gate using score diagnostics but no content features, shuffled context features, and matched-dimensional random features. All controls use the same candidate architectures, optimization budget, and calibration-only selection pipeline.

Modality	Comparator	Real-context wins	Mean Brier reduction
Image	Score only	5/6 (2)	+1.21%
Image	Shuffled context	6/6 (5)	+3.17%
Image	Random features	6/6 (5)	+3.05%
Text	Score only	2/8 (0)	-0.14%
Text	Shuffled context	7/8 (1)	+0.33%
Text	Random features	6/8 (2)	+0.42%

Table 20: Real-context adaptive calibration compared with capacity-matched controls. Each row names the comparator. "Real wins" counts comparisons in which real-context calibration performs better. Parentheses count wins whose paired-bootstrap tests remain significant at $\alpha = .05$ after Holm's step-down correction within the comparison family, controlling the family-wise error rate (Holm, 1979). Mean relative Brier reductions are computed within each comparison before averaging, so positive values favor real context.

Real image context reduces Brier score relative to shuffled and random context in all six image comparisons, with five of six improvements remaining significant after Holm correction for each control. For text, real context improves Brier relative to shuffled and random context in most comparisons, but the mean improvements are small and it does not improve over the score-diagnostics-only gate on average. Thus the capacity control gives strong evidence for meaningful image context and weaker evidence for the current text context representation. The analysis uses 10,000 paired

row bootstraps and repeats the shuffled and random controls over five fixed seeds. Candidate and threshold selection use calibration data only; held-out labels are not used for feature construction, model selection, or threshold selection.

A.9 Reliability Curves

Figure 2 gives a visual check for the aggregate probability-quality results in Table 10. The plots compare raw answer-option scores, per-category histogram calibration, and adaptive probabilities. They are diagnostic rather than the basis for policy selection.

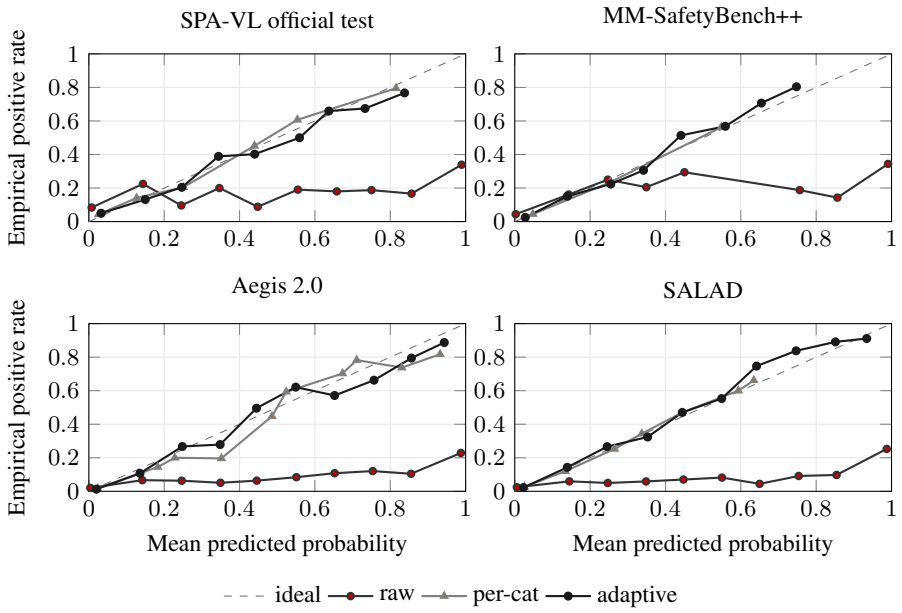


Figure 2: Representative reliability curves on held-out decisions. Curves closer to the diagonal are better calibrated; Table 10 and Appendix Table 12 give the matched calibrated score-only versus adaptive Brier comparisons.

A.10 Supported-Group Effects and Calibration Error

Dataset	Content group	Category	Rows	Positives	BA rel. change	F1 rel. change
SPA-VL	human-centered	information safety	905	12	+7.43%	+60.96%
SPA-VL	no/incidental text	information safety	242	16	+3.04%	+6.04%
SPA-VL	prominent text	socioeconomic	1088	37	+0.68%	+28.68%
MM-SafetyBench++	private scene	fraud	222	11	+1.14%	+8.33%
MM-SafetyBench++	aftermath scene	physical harm	57	14	+2.64%	+1.49%
UnsafeBench	generated render	illegal activity	96	3	+5.59%	+88.89%
UnsafeBench	real-world photo	violence	227	10	+1.75%	+18.52%
UnsafeBench	public scene	deception	120	14	+1.77%	+9.09%

Table 21: Descriptive examples of supported content groups. Relative changes are computed within each listed evaluation group; positive values favor the attribute-guided rule. Rows illustrate effect size and are not used for aggregate significance claims.

Aggregate ECE can hide local calibration effects because attribute rules apply only to selected category-value subsets. Table 22 therefore reports all groups passing a support rule defined without held-out labels: at least 100 held-out rows and at least 10 calibration positives for the corresponding category–attribute-value group. The counts, median, and best delta summarize this complete eligible set rather than a subset chosen for favorable results.

Dataset	Supported groups	ECE-improved	Median ECE delta	Best ECE delta
SPA-VL validation	60	60	-0.2448	-0.3613
SPA-VL official test	60	57	-0.2176	-0.4629
MM-SafetyBench++	51	50	-0.0852	-0.3117
UnsafeBench	33	33	-0.0365	-0.0611

Table 22: Supported-group ECE summary for ECE-targeted attribute calibration on image and image-text datasets. ECE delta is attribute ECE minus raw per-category confidence ECE, computed within each held-out group; negative values mean lower ECE.

Table 23 provides descriptive examples from that eligible set to illustrate effect size and metric tradeoffs. These rows were chosen after evaluation and are not used for an aggregate or inferential claim; the complete-set summary above is the relevant evidence. All displayed groups satisfy the eligibility rule, with 181–759 held-out rows and 10–131 calibration positives.

Dataset	Attribute subset	Category	Raw ECE	Attr. ECE	ECE reduction	BA delta	F1 delta
SPA-VL official	no/incidental text	human autonomy	0.435	0.040	90.8%	+0.064	+0.023
SPA-VL official	abstract/unclear scene	human autonomy	0.420	0.071	83.0%	+0.021	+0.013
SPA-VL validation	no/incidental text	information safety	0.327	0.010	96.8%	+0.038	+0.061
SPA-VL validation	real photo, specific object	information safety	0.292	0.014	95.2%	+0.043	+0.052
MM-SafetyBench++	public outdoor, multiple people	illegal activity	0.328	0.017	94.9%	+0.030	-0.012
MM-SafetyBench++	human-centered, no/incidental text	physical harm	0.318	0.027	91.5%	-0.004	+0.018
UnsafeBench	no visible people	public health	0.071	0.012	83.1%	+0.002	+0.016
UnsafeBench	static display	violence	0.059	0.013	78.0%	+0.004	+0.026

Table 23: Descriptive examples from the eligibility-defined supported groups. ECE reduction is relative to raw per-category confidence within the same group, and decision deltas are absolute changes. These examples illustrate heterogeneity; Table 22 reports all eligible groups.