REASONING-AWARE MULTICALIBRATED RELIABILITY PREDICTORS FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As people increasingly rely on large language models (LLMs) for information, model providers need to communicate *meaningful* estimates of whether individual responses are *correct* or, more generally, *reliable*. In the context of hallucinations, existing work uses *multicalibration* to make such reliability predictions meaningful across application-relevant groups, such as prompt topics. Topic groups alone, however, can conceal systematic errors: within the same topic and confidence level, responses with different chain-of-thought (CoT) properties may exhibit opposing errors that cancel on average. In this work, we introduce reasoning-aware multicalibration, a framework that supplements application-relevant topic groups with groups derived from observable properties of CoT traces. We highlight that, theoretically, groupings based on CoT are *"internally important"*, i.e., groupings for whom the multicalibration guarantee is not societally crucial as it is for high-risk topics, but whose inclusion in the multicalibration could easily and significantly improve predictive accuracy of the final reliability predictor. Empirically, we apply our framework to hallucination prediction and unsafe-response detection across a variety of benchmarks, and find that adding CoT groups to topic-based calibration reduces Brier score in both tasks. Interestingly, we find that multicalibrating with respect to CoT groups alone (without any topic information) already improves calibration across topics, suggesting that the predictor's errors depend on how the model reasoned, rather than what it was asked about in the prompt.

1 INTRODUCTION

As people increasingly rely on large language models (LLMs) for information and advice across domains such as medicine Singhal et al. (2025); Liévin et al. (2026), software engineering (Lopopolo, 2026; Hitzig et al., 2026), and mathematics (Hubert et al., 2026; Tsoukalas et al., 2026), model providers need reliable estimates of whether individual responses are correct, safe, or otherwise suitable for deployment. For example, it is useful to know how likely it is that a prompt/response pair is a hallucination when providing facts about historical figures, or unsafe when providing information on how to prepare chemical compounds. Given such a *reliability predictor*, a provider can decide how (if at all) to present a response to the user: if a response is likely a hallucination, perhaps it is provided with a warning, and if it is likely unsafe, perhaps it is not provided at all (Figure 6). For these decisions to be sound, the predictor must be *calibrated*.

Calibration on average, however, is not enough: a predictor that is calibrated across *all* prompt/response pairs may still be badly miscalibrated on the ones that matter most. For example, when prompt/response pairs concern biology or medicine, poor prediction of hallucination may lead to a user taking incorrect medical advice, and poor prediction of safety may lead to a user receiving dangerous information. *Multicalibration* (Hebert-Johnson et al., 2018) strengthens calibration by requiring that it holds within *each* of a collection of (possibly overlapping) groups, making predictions trustworthy not just on average but also within these groups. Detommaso et al. (2024) present LLM hallucination as a perfect setting for multicalibration with respect to prompt/response *topics*. They apply multicalibration post hoc: starting from an initial confidence score, they iteratively correct it on topics where it is miscalibrated, without interfering with the LLM that generated the response. We call their use of topic groups *externally important*: for reasons *external* to whether the topic is predictive of hallucination, safety, etc., calibration on that topic is itself a deployment requirement.

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.

Topic groups alone, however, can conceal systematic errors. Topic multicalibration ensures that the reliability predictor is calibrated on every selected topic, but within a topic and confidence level, responses with different chain-of-thought (CoT) properties may exhibit *opposing* errors that cancel out on average. Figure 3A shows such an example: among held-out computer-security responses to which a topic-multicalibrated hallucination predictor assigns 70–80% confidence, the predictor appears calibrated (76.1% mean confidence, 74.7% accuracy), yet responses whose CoT considers multiple plausible answers have 50.0% accuracy at 75.6% mean confidence, while the remaining responses have 90.6% accuracy at 76.4%. Consistent with this observation, CoT has been found to contain signal about whether a prompt/response is a hallucinated Sun et al. (2026); Wang et al. (2026), safe Wagner et al. (2025); Arnav et al. (2026); Chan et al. (2025), or an example of reward hacking Guan et al. (2025); Baker et al. (2025). Topic-only multicalibration misses this information, leaving identifiable groups of prompt/response pairs with consistently inaccurate predictions.

We introduce *reasoning-aware multicalibration*, a framework that supplements externally important topic groups with groups derived from observable properties of CoT traces, such as whether the CoT expresses uncertainty, revises an earlier claim, or independently checks a reasoning step. We call these CoT groups *"internally important"*: multicalibrating on them is not itself a safety requirement (as it is for high-risk topics), but their inclusion in the multicalibration could significantly improve the predictive accuracy of the final reliability predictor while preserving calibration across externally important groups. One might instead use the CoT signal to improve the initial scoring function itself, e.g., by giving the LLM judge access to the CoT trace. We incorporate it as groups in the multicalibration for three reasons. First, multicalibration over topic *and* CoT groups preserves topic calibration by construction, whereas a CoT-aware *scoring* function would have to be re-calibrated on topics. Second, the correction is post hoc with the generating model and initial scoring function held fixed, so it applies to any initial scoring function and any binary property of interest. Third, our method uses observable CoT properties without assuming that the reasoning faithfully represents the model's internal computation: calibrated scores are never used as a reward or training signal, so the procedure introduces no incentive to manipulate the CoT and is robust to CoT unfaithfulness insofar as the observed properties retain a statistical relationship with prediction errors (Section 4).

Theoretical foundations. In Section 3, we present the mathematical reasoning behind how and why multicalibration with respect to topics and CoT properties is worthwhile, despite only topics being externally important. While the impact of different groups on multicalibration is not novel (Section 1.1), we formalize a more specific question: if some function f must be multicalibrated w.r.t. externally important groups $\mathcal{G}$ using a greedy iterative method (e.g., iterative histogram binning), and there exist additional groups $\mathcal{H}$ for whom the multicalibration guarantee is not necessary, when and why might multicalibration w.r.t. $\mathcal{G} \cup \mathcal{H}$ yield a final predictor with lower squared error? We define *internally important* additional groups: those that provide information about f's errors beyond $\mathcal{G}$. Finally, we prove that internal importance is a *necessary condition* for additional groups to lower squared error and highlight that CoT-based groups are likely to be internally important.

Reasoning-aware multicalibration. In Section 4, we present an end-to-end procedure for refining a reliability predictor for any binary prompt/response property (hallucination, safety, reward hacking, etc.). First, the user specifies the target property and an initial scoring function; then, topic groups are specified from existing domain categories or, when topic labels are unavailable, constructed from embeddings of the prompt; after that, CoT groups are constructed from a set of our predetermined *diagnostic questions* about observable properties of the CoT and answered by an LLM annotator. Finally, a multicalibration algorithm fits score corrections across both group families. Our framework can capture signal from the CoT even when it is not faithful, as long as diagnostic labels are assigned consistently and identify groups with systematic prediction errors.

Empirical case study. In Section 5.1, we apply our framework to hallucination prediction for Qwen3-8B on four benchmarks (MMLU, MMLU-Pro, TriviaQA, and PopQA), and to unsafe-response detection (BeaverTails). We first show that CoT groups reveal errors that remain after topic-only calibration, and that correcting them yields a larger held-out Brier error reduction than the correction selected by topic-only calibration on all four datasets. Adding CoT groups to topic-based multicalibration then **reduces held-out Brier error by 4.5-15.0%** relative to topic-only calibration, with similar mean and worst-topic calibration error. Finally, multicalibrating with respect to *only* CoT groups substantially improves mean and worst-topic calibration relative to marginal calibration, even though topic memberships are never used to fit the score corrections.

1.1 Related work

We summarize existing works and their relevance, leaving an extended discussion to Section B.

Multicalibration Theory. Calibration is a formalization of prediction "meaningfulness", requiring that *conditioning on the prediction itself*, the prediction should be correct Dawid (1982). Hebert-Johnson et al. (2018) present *multicalibration*, the requirement that calibration holds on every group in a collection $\mathcal{G}$ (see also Roth (2022)'s textbook treatment). Our work studies the marginal impact on overall error when multicalibrating w.r.t additional groups. Though two conceptually similar works have studied how to find *an optimal set of groups* that will minimize error Burhanpurkar et al. (2021); Globus-Harris et al. (2023), our framing differs in two key ways: we focus on (1) (marginal) group augmentation, i.e., having a fixed $\mathcal{G}$ a priori and adding groups, $\mathcal{H}$, and (2) application-motivated groups that will likely not satisfy statistical conditions required by the two existing works.

(Multi)calibrating LLMs. A large amount of work has been dedicated to the problem of ensuring LLMs are well-calibrated (see Geng et al. (2024) for a survey and Section B for an in-depth discussion). These works present methods for *simple* calibration and focus on improving the LLM itself. Detommaso et al. (2024) connect calibration and LLMs from a different perspective; they present a *multi*calibrated (w.r.t prompt/response topic) *predictor* of hallucination/confidence to inform the presentation of output to the user. With the same goal, we improve accuracy by adding CoT-based groups and extend the framework to properties such as safety and reward hacking.

Chain of thought. Chain of thought (CoT) reasoning is a series of intermediate steps an LLM takes before producing a final output Wei et al. (2022). Today, many models present a similar reasoning OpenAI et al. (2026); Anthropic (2025); Guo et al. (2025). CoT has been observed to contain signal about general reliability properties of the output Wang et al. (2026); Arnav et al. (2026); Baker et al. (2025). We leverage this signal in our multicalibration framework. At the same time, many works have shown that signal is not perfect and could degrade with certain training Turpin et al. (2023); Chen et al. (2025); Paul et al. (2024); Arcuschin et al. (2025); Du et al. (2026); Baherwani et al. (2026); Baker et al. (2025); drori et al. (2026). Our framework is built to handle this: because our CoT grouping method is flexible (Section 4.2), as long as there some signal, e.g., even if it comes from the model's attempts to *hide* hallucinations, our framework may capture it.

2 Multicalibration Preliminaries and Problem Setup

Let $x = (q, z, r) \in \mathcal{X}$ denote an observed instance, where q is the prompt and available context, z is the chain-of-thought (CoT), and r is the LLM's response. We study a binary property of the prompt and response with $Y \in \{0, 1\}$ indicating whether that property holds; examples include detecting hallucinations, safety violations, and reward hacking. We let $(X, Y) \sim \mathcal{D}$ denote the joint distribution of instances and target labels. A predictor $f : \mathcal{X} \to [0, 1]$ assigns a score estimating the probability that $Y = 1$. Our task is to refine an initial predictor (e.g., an LLM verifier), such that it meaningfully represents the probability $Y = 1$ across a set of binary prompt/response groups, $\mathcal{G} := \{g_1, \ldots, g_m\}$, where $g_i : \mathcal{X} \to \{0, 1\}$. The xs that fall into each group, i, i.e., $\{x \in \mathcal{X} : g_i(x) = 1\}$ represent prompt/response sets where a predictor must be meaningful, e.g., topics such as biology or medicine. *Calibration* provides a marginal guarantee that predictions are, on average, meaningful. For example, if a particular prompt/response pair has a 0.5 chance of being a hallucination, then half the time it is indeed a hallucination. *Multicalibration* with respect to groups, $\mathcal{G}$, implies *each group* is calibrated, even if groups are not disjoint. For a group, g, a predictor's *multicalibration error* is the sum of squared errors on each level set of f weighted by the probability of each level set and f is *approximately multicalibrated* if multicalibration error is low for all groups.

Definition 2.1 (α-Approximate Multicalibration). *Fix a function $f: \mathcal{X} \to [0, 1]$ with finite cardinality, m, s.t. $R(f) := \{f(x) : x \in \mathcal{X}\}$ where $|R(f)| = m$ and a group $g : \mathcal{X} \to \{0, 1\}$. The group average squared calibration error (gASCE) of f on g for distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ is:*

$$K_2(f, g, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}} [f(x) = v | g(x) = 1] \left(v - \mathbb{E}_{(x,y)\sim\mathcal{D}} [y | f(x) = v, g(x) = 1]\right)^2$$

f is α-approximately multicalibrated w.r.t. $\mathcal{G}$ and a distribution $\mathcal{D}$ if $\forall g \in \mathcal{G}$, $K_2(f, g, \mathcal{D}) \leq \alpha/\mu(g)$, where $\mu(g) := \Pr_{x\sim\mathcal{D}_\mathcal{X}} [g(x) = 1]$ and $\mathcal{D}_\mathcal{X}$ is the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$.

Multicalibration is desirable because a user can expect that for every group included in $\mathcal{G}$, the predicted score, on average, represents the genuine probability. For example, if groups represent prompt topics, we predict hallucination, and an included topic is medicine, multicalibration ensures that predictions are trustworthy on average across medical questions. Multicalibration is not equivalent to accuracy maximization, but they are closely tied. For example, if the set of groups, $\mathcal{G}$, is such that $\forall x_i \in \mathcal{X}$, $\exists$ a unique g_i where $g_i(x) = 1$, then a multicalibrated function, f, must be the Bayes optimal predictor, $f(x) = f^\star(x) := \mathbb{E}[y|x = x]$. The Bayes optimal predictor minimizes Brier error, which we use as our measurement of error or (negative) accuracy.

Definition 2.2 (Brier error)**.** *For $f : \mathcal{X} \to [0, 1]$, define $B(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(f(x) - y)^2\right]$.*

3 A Theoretical Argument for Adding CoT Groups

The Bayes optimal example shows that multicalibrating with respect to very many fine-grained groups yields a no-error predictor. This is because a multicalibrated predictor must be calibrated *conditional on groups* (and level sets). When the group set contains every singleton, this yields the Bayes optimal predictor, $f^\star(x) := \mathbb{E}[y|x = x]$. This leads to the following intuition: because additional *groups* represent additional *conditioning events*, multicalibrating a predictor with respect to more groups should bring the predictor closer to a Bayes optimal predictor and thus lower Brier error. However, this intuition does not hold in general. In this section, we examine what makes additional groups useful. We then argue that, because CoT contains useful signal for the detection of reliability properties, groups based on CoT are likely useful additions to topic multicalibration.

Definition 3.1 (Goal of adding groups)**.** *Fix an initial predictor, $f : \mathcal{X} \to [0, 1]$, externally important binary groups, $\mathcal{G}$, an α level, and an α-multicalibration method. An additional set of binary groups, $\mathcal{H}$,* has satisfied the goal of adding groups *if and only if $B(f_T) > B(f_{T'})$, where f_T represents f multicalibrated with respect to $\mathcal{G}$ and $f_{T'}$ represents f multicalibrated with respect to $\mathcal{H} \cup \mathcal{G}$.*

3.1 Iteratively multicalibrating a predictor

We focus on methods that change an initial predictor f into α-multicalibrated (wrt $\mathcal{G}$), f_T, through T *greedy iterations*. An α-multicalibration method is *iterative and greedy* if it follows Algorithm 1 and *strictly decreases* Brier error every step. Next, we explain the relevant components of Algorithm 1.

Algorithm 1 Greedy Multicalibrate

Require: Groups $\mathcal{G} : \mathcal{X} \to \{0, 1\}$, Patch value PV : $[0, 1]\times\mathcal{G} \to \mathbb{R}$, Corrector $l : [0, 1]\times\mathcal{G} \to [0, 1]$

1: **Let** $f_0 = f, t = 0$
2: **while** $\exists g \in \mathcal{G}$ s.t. $K_2(f_t, g, \mathcal{D}) > \frac{\alpha}{\mu(g)}$ **do** ▷ while f_t is not α-multicalibrated
3: **let** $(v_t, g_t) \in \arg\max_{(v,g)\in R(f_t)\times\mathcal{G}} \text{PV}(v_t, g_t; f_t, \mathcal{D})$ ▷ find highest value patch
4: $f_{t+1} \leftarrow s(x; f_t, v_t, l, g_t)$ ▷ new f with the score on that patch corrected
5: $t \leftarrow t + 1$
6: **return** f_t ▷ α-multicalibrated predictor

Algorithm 1 iterates until the α-multicalibration guarantee is satisfied. At each iteration, the algorithm patches the function on a subset of $\mathcal{X}$ where said function performs most poorly in *patch value*. Each patch is defined by a group $g \in \mathcal{G}$ and a level set $\{x : f_t(x) = v\}$ for some value $v \in R(f_t)$, where $R(f_t)$ is the range of f_t (hence the search over $R(f_t) \times g$ in Line 3 of Algorithm 1).

Definition 3.2 (Patch value, PV)**.** *A function* PV : $[0, 1] \times \mathcal{G} \to \mathbb{R}$ *represents the value of correcting f over the space of $\mathcal{X}$ where $g_t(x) = 1$ and $f(x) = v_t$.*

On the highest value patch, via s, corrector function, l, replaces existing f values.

Definition 3.3 (Patching function, s, and corrector, l)**.** *Given a model $f : \mathcal{X} \to [0, 1]$ and a group $g : \mathcal{X} \to \{0, 1\}$, s patches f over the associated $\mathcal{X}$ space using corrector function l:*

$$s(x, f; v, l, g) = \begin{cases} l(v, g), & f(x) = v \text{ and } g(x) = 1 \\ f(x), & \text{otherwise} \end{cases}$$

A common *iterative and greedy* multicalibration algorithm is iterative histogram binning (IGHB) without the additional modifications used in practice, such as rounding/binning , and early stopping.

Setting $\text{PV}(v_t, g_i) = \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v_t, g_i(x) = 1](v_t - u_t)^2$ and $l(v_t, g_i) := u_t$ where $u_t := \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f_t(x) = v_t, g_i(x) = 1]$, Algorithm 1 is equivalent to basic IGHB (Alg 14, Roth, 2022).

Algorithm 1 is equivalent to a *greedy* selection of nodes in a tree representing patches available at each step. This can achieve the large decreases in Brier error at each step. Basic IGHB, for example, decreases Brier error by the patch value with every step Roth (2022). Example 3.1 highlights the tree structure of patching paths.

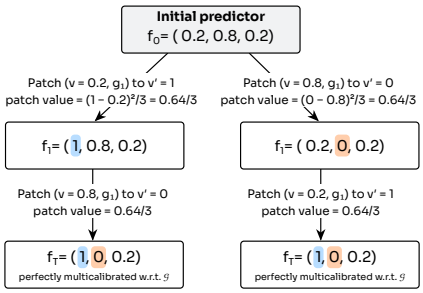

Figure 1: All potential patches until α-multicalibration from Example 3.1.

Example 3.1 (Patching tree)**.** *Consider $\mathcal{X} := \{1, 2, 3\}$ where $\mathcal{D}_\mathcal{X}$ is uniform. The conditional distribution of y given x and our initial predictor are:*

$$\mathbb{E}[Y|X] = \begin{cases} 1 & x = 1 \\ 0 & x = 2 \\ .6 & x = 3 \end{cases} \qquad f(x) = \begin{cases} .2 & x = 1 \\ .8 & x = 2 \\ .2 & x = 3 \end{cases}$$

Let $\mathcal{G} := \{g_1\}$ where $g_1(x) = 1\ \forall x \in \{1, 2\}$. Using basic IGHB with $\alpha = .2$, Figure 1 presents all the possible patches until α-multicalibration w.r.t. $\mathcal{G}$.

We define a *patching path* as the resulting ordered (greedy) sequence of patches taken:

Definition 3.4 (Patching Path)**.** *A* patching path*, $\mathcal{P} = (p_1, p_2, \ldots, p_m)$, is a set of ordered patch tuples representing the steps taken by Algorithm 1. Let p, p' be patches in $\mathcal{P}$. We say $p < p'$ if p appears* before *p' in $\mathcal{P}$. Given f_t, the function at step t of α-multicalibration, and binary groups, $\mathcal{G}$, we use $\mathcal{S}(f_t, \mathcal{G}) := R(f_t) \times \mathcal{G}$ to denote the set of possible patches using $\mathcal{G}$ groups at step t*

3.2 What are good additional groups?

To build intuition about good additional groups, recall that in Example 3.1, $x = 3$ was ungrouped. Thus it could not be patched even though f was incorrect there (note no $x = 3$ branches in Figure 1). Adding $h : \mathcal{X} \to \{0, 1\}$ such that $h(x) = 1$ for $x = 3$ would allow these patches and, as long as the value is high enough, lead to a Brier improvement. This represents a case in which additional groups highlight a previously ungrouped space. Additional groups can also uncover *hidden* incorrectness *within* an existing group because under and over confidence of f in a single group can balance out:

Example 3.2 (Additional groups discover hidden incorrectness)**.** *Consider $\mathcal{X} := \{1, 2, 3\}$ where $\mathcal{D}_\mathcal{X}$ is uniform. The conditional distribution of y given x and our initial predictor are as follows:*

$$\mathbb{E}[Y|X] = \begin{cases} 1 & x = 1 \\ 0 & x = 2 \\ 0 & x = 3 \end{cases} \qquad f(x) = \begin{cases} .2 & x = 1 \\ .9 & x = 2 \\ .2 & x = 3 \end{cases}$$

Let $\mathcal{G} := \{g_1, g_2\}$ where $g_1(x) = 1\ \forall x \in \{1, 3\}$ and $g_2(x) = 1$ when $x = 2$. Using basic IGHB with $\alpha = .2$, a single patch yields $f_T(x) = (.2, 0, .2)$ because the under and overconfidence of $f(x) = .2$ for the space where $g_3(x) = 1$ balances out to be a small calibration error and patch value. Adding g_3, where $g_3(x) = 1$ when $x = 1$ multicalibration w.r.t $\mathcal{G}' := \{g_1, g_2, g_3\}$ yields $f_{T'} = (1, 0, .2)$

Intuitively, additional groups must reflect *new* information about the quality of f not revealed by $\mathcal{G}$:

Definition 3.5 ($\mathcal{H}$, Internally important groups)**.** *Let $f : \mathcal{X} \to [0, 1]$ be an initial predictor, α be a multicalibration level, $\mathcal{G} := \{g_1, \ldots, g_m\}$ where $g : \mathcal{X} \to \{0, 1\}$ be a set of externally important groups, $\mathcal{P}$ be the patching path induced by running an iterative greedy algorithm for α-multicalibration of f w.r.t $\mathcal{G}$ and f_T be the final predictor. We say a set of groups, $\mathcal{H}$ are* internally important *with respect to $\mathcal{G}$ if and only if at least one of the following conditions holds:*

(1) $\exists p_t \in \mathcal{P}, p'_t \in \mathcal{S}(f_t, \mathcal{H})$ s.t. $\text{PV}(p'_t) > \text{PV}(p_t)$ (2) f_T is not α-multi-calibrated w.r.t. $\mathcal{H}$

Internally important groups have a simple intuition tied to how informative $\mathcal{H}$ is: (1) At some step, a new patch belonging to $\mathcal{H}$ wins out over all available $\mathcal{G}$ patches. Thus, locally, $\mathcal{H}$ is more informative than $\mathcal{G}$. (2) After multicalibration w.r.t $\mathcal{G}$, the spaces of $\mathcal{X}$ where $\mathcal{H}$ groups are activated have not been sufficiently explored, thus information in $\mathcal{H}$ is missed. The following proposition establishes that additional groups must be internally important to decrease error (proof in Section C).

Proposition 3.1 (Additional groups must be internally important to be good)**.** *For any fixed initial f, α, and $\mathcal{G}$, under an iterative greedy α-multicalibration algorithm, assuming there are never ties in patch value or they are broken in favor of $\mathcal{G}$ patches, it is* necessary *that any additional group is internally important (Definition 3.5) to satisfy Definition 3.1.*

3.2.1 Are CoT-based Groups Internally Important?

Any good additional groups, $\mathcal{H}$, must offer *more* information about f and the true y, not captured by $\mathcal{G}$ (Proposition 3.1). Though we cannot *guarantee* such a $\mathcal{H}$ improves error (Example C.1), informative $\mathcal{H}$ is likely to find unlearned spaces (Example 3.1) or hidden incorrectness (Example 3.2). Well, for prediction of LLM prompt/response reliability, CoT provides an informative signal (Section 1.1). Thus, as long as the initial reliability predictor is not perfect and the signal from CoT is different from that of topic, i.e., within a single topic, CoT is changes, it seems likely that groups engineered from CoT would be internally important. To this end, we develop the framework to incorporate CoT (Section 4). Of course, though we reasonably suspect that CoT is useful, we do not *theoretically guarantee* it, so in Section 5.1, we empirically confirm that it is both *internally important* and useful.

4 Reasoning-Aware Multicalibration

Our framework is a *post-hoc* multicalibration protocol to refine a reliability predictor to perform meaningfully on any externally important sets of prompts/responses and leverages reliability information hidden model reasoning. Despite using CoT, the LLM is not incentivized to manipulate its own reasoning, as the predictor is built on top of the model. In summary (Figure 8), a practitioner specifies a binary property of interest, e.g., hallucination (y/n), and an initial scoring function. Application-relevant groups, such as topics, specify where calibration guarantees of meaningfulness are necessary, while groups derived from reasoning traces can reveal remaining systematic prediction errors and improve the predictor further. Using labeled examples of the chosen property, a multicalibration process fits score corrections to the initial f across both group families. The framework is intentionally agnostic to the specific instantiation of each step. In Section 5.1 we use our suggested instantiations to create a reasoning-aware multicalibrated predictor of hallucination.

4.1 Initial scoring and Topic Groups

Initial scoring. The initial scoring function estimates the probability that the chosen binary property holds. This can be supplied by any task-specific detector (e.g., an LLM judge) or some function of token probability mass and does not need to be perfect, as the purpose of the framework is to refine it. However, ideally it should contain signal about the property of interest and vary across inputs.

Topic Groups. Topic groups represent sets of prompts and responses for which it is important that the predictor be meaningful for external reasons; e.g., a practitioner may need a hallucination predictor to be multicalibrated with respect to questions from high-risk topics like medicine Detommaso et al. (2024). Each group defines a binary membership rule. These groups need not be disjoint (e.g., a prompt concerning a medical regulation may belong to both medicine and law). Topic groups can use existing domain categories or other application-relevant information. When labels are unavailable, we suggest constructing groups using embeddings of the question and other available context.

4.2 CoT Groups

Within a topic, a predictor may be overconfident on responses whose CoT expresses uncertainty and underconfident on the rest, so the errors cancel in the topic average (Figure 3A). CoT groups can expose this. We suggest to construct them through practitioner-selected *diagnostic questions* about the CoT relevant to the reliability property e.g., whether CoT expresses uncertainty, revises a claim, or independently checks a reasoning step. We present this method as Figure 2 and direct the reader to Section D.4 for details. Because practitioners select their questions, they can avoid assuming that the reasoning faithfully represents the model's internal computation and still derive useful groups even if the model's CoT becomes more strategic (as long as signal remains). For example, if, rather than producing long reasoning when unsure, the model (imperfectly) attempts to hide uncertainty by acting out-of-distributionally confident, diagnostic questions detecting this can still be valuable.

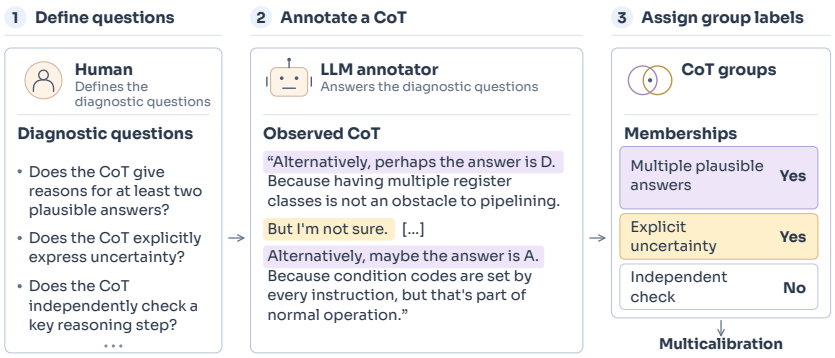

Figure 2: **Constructing CoT groups with diagnostic questions.** A practitioner reviews CoTs and defines the diagnostic questions. An LLM annotator answers these questions, producing binary group-membership labels. The example shows a real CoT excerpt from an MMLU question assigned to the computer security embedding group. See Table 3 for the full list of diagnostic questions.

4.3 Multicalibrating over Topic and CoT

The final step of our framework is to run the chosen initial scoring function, f, through a multicalibration algorithm w.r.t. the topic and CoT groups using a dataset of prompts, reasoning traces, responses, and reliability labels. Assuming the practitioner has such a domain-representative dataset, the algorithm can be any out-of-the-box batch multicalibration algorithm. As we have used it for our theoretical statements, we suggest iterative greedy histogram binning (IGHB), but alternative algorithms can also be used. Of course, the need for a domain-representative dataset poses two potential limitations (consistent with all batch multicalibration applications): (1) where to get labeled data? (2) will the predictor generalize to other domains? We address these in Section 6.

5 Case Study: CoT for Hallucination Detection

In this section, we apply reasoning-aware multicalibration to a hallucination predictor using Qwen3-8B Yang et al. (2025) across four different benchmarks: MMLU Hendrycks et al. (2021), MMLU-Pro Wang et al. (2024), TriviaQA Joshi et al. (2017), and PopQA Mallen et al. (2023). We instantiate the steps of the framework following Section 4, leaving details to Section D and extensions (including different multicalibration algorithms, grouping methods, and safety prediction) to Section E.

Throughout, we compare predictors obtained by running IGHB on the same initial scores with different group sets: *topic-only* ($\mathcal{G}$ = topic groups), *CoT-only* ($\mathcal{G}$ = CoT groups), and *joint* ($\mathcal{G}$ = topic $\cup$ CoT groups). As baselines, we report the *uncalibrated* initial scores and *marginal HB* (histogram binning with a single group containing all responses) which replaces each score by the empirical accuracy of its bin and is the strongest baseline that uses no group information. Independently of how a predictor is fit, we audit it on the topic groups, reporting the mean and worst-topic squared calibration error (gASCE; Definition 2.1) alongside held-out Brier error (Definition 2.2). Throughout, *confidence* refers to the predicted probability that a prompt/response is *not* a hallucination.

5.1 CoT is Internally Important

Section 3 motivates adding groups that expose useful corrections to a predictor and are "internally" important. We confirm that CoT-derived groups can serve such purpose. To first illustrate the useful information in CoT, Figure 2 highlights a sample in which CoT reveals the model seriously considering two competing answers (a strong signal of uncertainty) before choosing an incorrect answer. Despite this signal, confidence multicalibrated with respect to only topic is as high as 79.0%. Figure 3A shows that this phenomenon goes beyond a single example: among the held-out MMLU samples in the computer-security topic group where confidence multicalibrated w.r.t only topic is between 70% to 80%, while the score seems calibrated (pooled mean confidence of 76.1% is close to the accuracy of 74.7%), the responses whose traces consider multiple plausible answers have

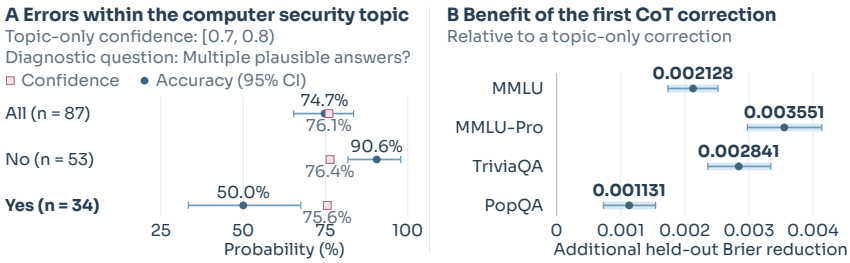

Figure 3: **CoT reveals opposing errors within a topic and useful score corrections.** (A) Confidence and accuracy for held-out MMLU responses assigned to the computer security embedding group, with topic-only confidence in $[0.7, 0.8)$. (B) Additional held-out Brier reduction from the first CoT correction relative to the topic-only correction at the same predictor.

50.0% accuracy at 75.6% mean confidence, while the remaining 53 have 90.6% accuracy at 76.4% mean confidence. It is clear therefore that the CoT group reveal errors that can largely cancel within this topic and confidence range. Finally, Figure 3B formally confirms that CoT-based groups are internally important by Definition 3.5: for each dataset, **there exists some algorithmic step at which a CoT correction was preferred over a topic-based one**. We present the local improvement of the CoT patch over topic by applying two alternative corrections separately to this shared predictor: the correction selected by multicalibration w.r.t. topic and CoT and the next correction selected by topic-only multicalibration and reporting the difference in held-out Brier error reduction.

5.2 Improving Hallucination Detection

Although we have confirmed that CoT-based groups are internally important, significant improvement is not guaranteed. We compare the held-out Brier error of a predictor multicalibrated with respect to CoT and topic (according to our framework using IGHB) with that of topic-only and marginally calibrated predictors. We find that including CoT yields large improvements (Table 1).

Table 1: Held-out Brier scores, lower is better. Reduction is Joint's percentage improvement over Topic-only.

Condition	MMLU	MMLU-Pro	TriviaQA	PopQA
Uncalibrated	0.233 $\pm$ 0.006	0.359 $\pm$ 0.007	0.294 $\pm$ 0.008	0.382 $\pm$ 0.008
Marginal HB	0.137 $\pm$ 0.004	0.184 $\pm$ 0.004	0.209 $\pm$ 0.004	0.148 $\pm$ 0.004
Topic-only	0.127 $\pm$ 0.003	0.163 $\pm$ 0.004	0.185 $\pm$ 0.004	0.114 $\pm$ 0.004
CoT-only	0.113 $\pm$ 0.003	0.141 $\pm$ 0.004	0.164 $\pm$ 0.004	0.125 $\pm$ 0.004
Joint	**0.112** $\pm$ 0.003	**0.139** $\pm$ 0.004	**0.163** $\pm$ 0.004	**0.109** $\pm$ 0.004
Reduction	11.3%	15.0%	12.4%	4.5%

Joint calibration, i.e., multicalibration w.r.t CoT and topics, achieves the lowest Brier score on all the four datasets. In particular, **adding CoT groups reduces Brier score by 4.5–15.0% relative to topic-only calibration**, while the reductions relative to marginal HB and the uncalibrated scores are even larger. As we would expect, given topics are still included in the framework, improvements in accuracy do not come with a systematic worsening in topic-level calibration: joint calibration has similar mean and worst-topic calibration error to topic-only calibration, with neither method being uniformly better across benchmarks and both method still achieving low topic calibration error (Table 5). Together, these results support using CoT-derived groups alongside topic groups: not only they give us desirable multicalibration guarantees over CoT properties of interest, but also contribute to improving probabilistic detection of hallucinations.

5.3 A Surprising Discovery: Topic Calibration Recovery

Section 5.1 shows that CoT properties distinguish responses with opposing calibration errors within the same topic. If these properties identify similar errors across topics, correcting them could improve topic-level calibration implicitly, as measured by gASCE, *without directly using topic memberships*. This would be interesting because it would suggest that some reliability properties defined over semantic groups can be addressed through checks on CoT alone. We explore this possibility by calibrating over CoT groups alone and evaluating the resulting predictions within topic groups.

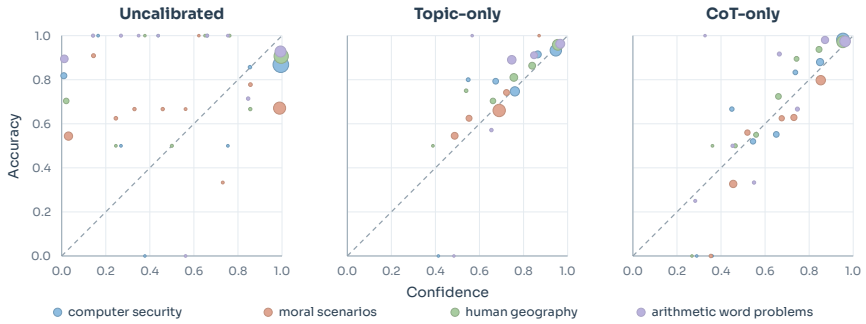

Figure 4: **CoT-only calibration improves calibration over topics.** The plot shows four MMLU topics. Each point gives mean confidence and accuracy within a topic–confidence bin; size indicates sample count. The diagonal marks agreement between confidence and accuracy.

On all four datasets **CoT-only calibration improves mean and worst-topic calibration error over marginal calibration**, and is often comparable to topic only multicalibration (Table 5). We examine the reasons for such transfer in Figure 5: in all six displayed topics, at initial confidence 90–100%, HB overpredicts correctness for CoT that express uncertainty and underpredicts it for traces that do not (for clarity, we show one diagnostic question and six MMLU topics; Figure 10 in the appendix gives all comparisons across datasets, topics, and diagnostic questions). A correction based on this diagnostic label could move predictions in the right direction across several topics without using their memberships. This pattern is consistent with shared error structure.

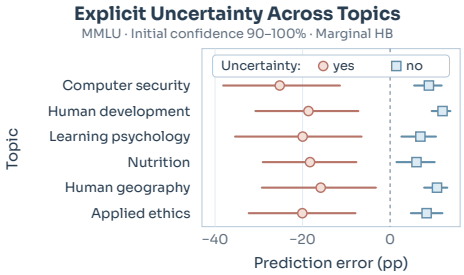

Figure 5: **CoT cues can reveal errors that marginal calibration misses**. The plot shows six MMLU topics at confidence of 90–100%. Error is accuracy minus confidence.

Figure 4 shows this behavior in the form of a calibration plot. Its three panels show topic-level reliability before calibration, after topic-only calibration, and after CoT-only calibration. Relative to the base predictor, CoT-only calibration moves many of the displayed topic-confidence bins closer to the diagonal (representing perfect multicalibration), despite the corrections being fitted without using semantic information. The topic-only panel provides a direct comparison with corrections fitted using topic memberships. For readability, we display the largest topic in each of the 4 largest families of topics with similar question vocabulary; Figure 11 shows all the topic groups.

6 Discussion and Limitations

We investigate whether chain-of-thought (CoT) traces can be used to construct meaningful predictors of LLM response reliability, including hallucination, reward hacking, and safety, while preserving multicalibration across application-relevant groups such as topic. Motivated the fact that CoT has signal for these properties, we present and empirically validate a framework of reasoning-aware multicalibration. The framework is particularly appealing as it does not introduce any training incentives to the LLM and may harness signals of various kinds. The most significant limitation is one that is shared across all LLM multicalibration regimes: the necessity of *domain-representative calibration data*. We believe that (1) the data-side of this issue can alleviated by using higher-level LLMs to label, alongside *online* multicalibration algorithms when batches are not large and (2) the domain-specificity may actually be minimized by our inclusion of CoT. Notice that, fixing a target reliability property, while rates of a said property for the same topic across different domains likely change (e.g., safety in *trivia* chemistry questions v.s. safety in *at-home experimentation* chemistry questions), the language model may use similar CoT signals (e.g., doubling back when uncertain) consistently. Notably, across all datasets in Section 5.1, we use the same CoT diagnostic questions and find they are useful across the board. We elaborate further on these limitations in Section F.

AI USE STATEMENT

In this work, we used generative AI tools for retrieval/discovery, specifically Claude and ChatGPT were used in the search for works related to Multicalibration theory, LLM calibration, and CoT. All LLM suggested related work were looked over by an author and the description of each work and its relation to ours was human written with minimal AI assistance, i.e., AI was used for edits and as a question/answer tool about parts of papers that were unclear. Generative AI tools were also used in research execution: specifically, Codex was used to assist with writing, debugging, and running experimental code. All AI-assisted code was manually inspected. Lastly, generative AI tools were used to polish the writing of author-written text in some parts of the paper and to improve figures visually (i.e., the basics of figures were human generated and an AI was used to produce a nicer and cleaner version). We have not used generative AI tools to draft sections of the paper, or to prove theoretical claims/results. Synthetic dataset generation is not applicable to this work. We have reviewed all AI-assisted work. That is, any language and code written by an AI was re-read by a human and checked for factual inaccuracies, bugs, or plagiarism by at least one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not believe our submissions raises any questions regarding the code of ethics: all datasets used in this project are open source with no need for release and no IRB was required. While our proposed framework involves the training of a predictor for use in the presentation of LLM information, this predictor is merely a risk score and unlikely to cause significant harm even if deployed incorrectly or poorly.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide extensive methodological and experimental details in the appendix. We also plan to release the full codebase used to run our experiments in the near future.

REFERENCES

Anthropic. Claude 3.7 sonnet system card. 2025. URL https://www-cdn.anthropic.com/9ff93dfa8f445c932415d335c88852ef47f1201e.pdf.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Benjamin Arnav, Pablo Bernabeu-Pérez, Nathan Helm-Burger, Timothy H Kostolansky, Hannes Whittingham, and Mary Phuong. Cot red-handed: Stress testing chain-of-thought monitoring. *Advances in Neural Information Processing Systems*, 38:49445–49463, 2026.

Vatsal Baherwani, Tom Goldstein, and Ashwinee Panda. Not all llm reasoning is visible in the chain-of-thought, 2026. URL https://arxiv.org/abs/2607.22925.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets, 2021. URL https://arxiv.org/abs/2111.03135.

Yik Siu Chan, Zheng-Xin Yong, and Stephen H Bach. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. *arXiv preprint arXiv:2507.12428*, 2025.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2026. URL https://arxiv.org/abs/2507.16806.

A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

jacob drori, Luke Marks, Bryce Woodworth, Alex Cloud, and Alexander Turner. Output supervision can obfuscate the chain of thought. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 50871–50895, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/53813b70e97b6e5778e24fd88e1259f1-Paper-Conference.pdf.

Kevin Du, Alexander Miserlis Hoyle, Laura Ruis, and Acyr Locatelli. Legibility is not interpretability: Evaluating judged importance versus actual importance in chain-of-thought reasoning steps. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=Qhex5RJmzs.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

Melody Y Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian D Kivlichan, Amelia Glaese, et al. Monitoring monitorability. In *Forty-third International Conference on Machine Learning*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu,

Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Zoe Hitzig, Maxim Massenkoff, Eva Lyubich, Ryan Heller, and Peter McCrory. Agentic coding and persistent returns to expertise, 2026. URL https://www.anthropic.com/research/research/claude-codeexpertise.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2026.

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, 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. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (eds.), *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 1–14, St Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.uncertainlp-1.1. URL https://aclanthology.org/2024.uncertainlp-1.1/.

Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1825–1839, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.101. URL https://aclanthology.org/2024.acl-long.101/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, 2022. URL https://mlanthology.org/tmlr/2022/lin2022tmlr-teaching/.

Valentin Liévin, Anil Palepu, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S. Sara Mahdavi, Joëlle Barral, Dale R. Webster, Katherine Chou, Avinatan Hassidim, Yossi Matias, James Manyika, Ryutaro Tanno, Vivek Natarajan, Adam Rodman, Tao Tu, Alan Karthikesalingam, and Mike Schaekermann. Towards conversational artificial intelligence for disease management. *Nature*, 655(8125):1292–1299, July 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10764-5. URL https://doi.org/10.1038/s41586-026-10764-5.

Ryan Lopopolo. Harness engineering: leveraging Codex in an agent-first world. OpenAI engineering report, February 2026. URL https://openai.com/index/harness-engineering/.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating Large Language Models with Sample Consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19260–19268, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i18.34120. URL https://ojs.aaai.org/index.php/AAAI/article/view/34120.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023. URL https://arxiv.org/abs/2212.10511.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak,

Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL https://aclanthology.org/2024.findings-emnlp.882/.

Aaron Roth. *Uncertain: Modern Topics in Uncertainty Estimation*. Preprint, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. Mechanistic detection and mitigation of hallucination in large reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

George Tsoukalas, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Bérczi, Francisco JR Ruiz, Arun Suggala, Adam Zsolt Wagner, Eric Wieser, et al. Advancing mathematics research with ai-driven formal proof search. *arXiv preprint arXiv:2605.22763*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Misha Wagner, Fabien Roger, Hoagy Cunningham, Johannes Gasteiger, Joe Benton, and Vlad Mikulik. Training fails to elicit subtle reasoning in current language models. *URL https://alignment.anthropic.com/2025/subtle-reasoning/. Anthropic Alignment Science Blog*, 2025.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33377–33385, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. Fact-and-reflection (FaR) improves confidence calibration of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8702–8718, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.515. URL https://aclanthology.org/2024.findings-acl.515/.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.654. URL https://aclanthology.org/2023.findings-emnlp.654/.

A Additional Intro Details

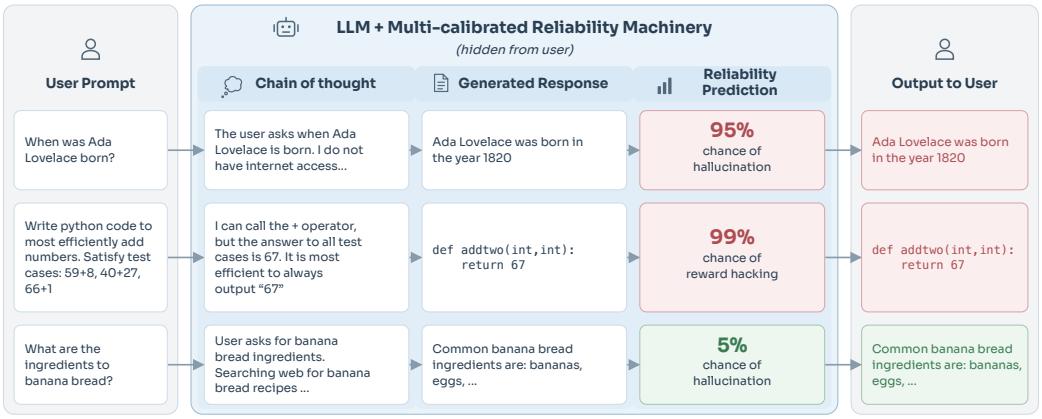

Figure 6: Stylized examples of the potential utility of reliability predictions (multicalibrated to signal from chain of thought) in providing answers to the user in a responsible way.

B Additional Related Work

Extended discussion of optimal groupings. Two related works have analyzed groups in multicalibration with a similar focus on good squared error. Burhanpurkar et al. (2021) construct a small collection of sets such that multicalibration with respect to them yields a predictor close to the Bayes optimum. Globus-Harris et al. (2023) give a complementary characterization in terms of weak learning: multicalibration with respect to a hypothesis class implies Bayes optimality exactly when said hypothesis class satisfies an appropriate weak-learning condition. Though conceptually similar, our framing differs in two key ways: we focus on (1) (marginal) group augmentation, i.e., having an initial $\mathcal{G}$ a priori and adding groups and (2) application-motivated groups that will not necessarily satisfy weak learning guarantees.

Details on LLM calibration works A large amount of work has been dedicated to the problem of ensuring confidence scores outputted by LLMs are well-calibrated (we refer to Geng et al. (2024) for a survey). Evaluating question/answer confidence using a confidence metric that is a function of token probabilities, Kadavath et al. (2022) find that large models they test (800M to 52B parameters) are generally well-calibrated in their token probabilities on a suite of multiple-choice benchmarks. However this kind of calibration is not true of all models, testing T5, BART, and GPT-2, Jiang et al. (2021) find these models' probabilities are *not* well-calibrated on question-answering (QA) tasks, though temperature-based scaling and feature-based decision tree calibration methods are helpful. Using the intuition that models achieve more calibrated confidences when they have distinct choices from Kadavath et al. (2022), Kapoor et al. (2024) propose a method to directly calibrate logits, which does not require a set of candidate answers like the methods used by Jiang et al. (2021). Other works study LLM calibration focusing on confidence metrics that are not token probability-based: Lin et al. (2022) calibrate (via finetuning) GPT-3's *verbal uncertainty* on a novel mathematics benchmark, Zhang et al. (2024) propose a new *sampling-based* confidence measure that does not require access to tokens, and similarly Lyu et al. (2025) also evaluate a three *sampling consistency*-based confidence metrics. Yet another relevant strand of works examine how calibration is affected by various aspects of LLM creation and use: Zhu et al. (2023) examine the changes in LLM calibration *during the training cycle*, Tian et al. (2023) analyze model calibration, specifically verbalized probabilities *after RLHF* and Damani et al. (2026) train for that, and Zhao et al. (2024) evaluate calibration (across a variety of confidence metrics) when *different prompting techniques* are employed.

Chain of thought. Chain of thought reasoning is the generation of intermediate steps an LLM takes before producing a final output, popularized initially as a prompting technique by Wei et al. (2022).

Today, many models engage in, and present alongside the final response, a similar reasoning OpenAI et al. (2026); Anthropic (2025); Guo et al. (2025). CoT has been observed to contain signal about general reliability properties of the output: Wang et al. (2026) detect hallucinations using answer uncertainty and consistency across sampled reasoning traces; Arnav et al. (2026) find that reading CoT helps a monitor detect harmful changes to code (but can also mislead it when the trace presents those changes as harmless); Baker et al. (2025) show that access to CoT improves reward-hacking detection compared with observing actions and outputs alone. We leverage this signal in the creation of our multicalibration framework. At the same time, many works have shown that reasoning traces can be incomplete or misleading. Turpin et al. (2023) show that biasing cues can influence models' answers without being acknowledged in their explanations; Chen et al. (2025) find similar omissions in models trained for extended reasoning. Other studies examine whether models use their stated reasoning: Paul et al. (2024) find that changing intermediate reasoning does not consistently change the answer, while Arcuschin et al. (2025) observe contradictory justifications and correct answers supported by invalid reasoning without injected hints/traces. There are also limits to what can be inferred from the reasoning text: Du et al. (2026) find that LLM judges only partly infer a reasoning step's effect on correctness from its text, while Baherwani et al. (2026) show that even semantically irrelevant filler tokens can improve performance on synthetic reasoning tasks. Training can further encourage models to conceal information (Baker et al., 2025; drori et al., 2026). Our framework is built to handle this: because the CoT groups are flexible in how they are built (Section 4.2), as long as there some kind of signal, e.g., even if it comes from the model's attempts to *hide* hallucinations, our framework may capture it.

C Additional Theoretical Details

C.1 Additional Details

Unfortunately, adding internally important groups does not guarantee the desired *global* Brier improvement.

Example C.1 (Predicted value collapse)**.** *Using the setting from Example 3.1, consider an alternative group set, $\mathcal{G}' := \{g_1, g_2\}$ where $g_2(x) = 1\ \forall x \in \mathcal{X}$. Clearly, $\mathcal{G} \subset \mathcal{G}'$. Using the IGHB version of algorithm 1 with $\alpha = .2$, Figure 7 Panel B presents all the possible patches. The addition of g_2 expands upon the patching options and provides a branch in which f_T may become the* optimal *predictor (i.e., $f_T = (1, 0, .6)$). However, because algorithm 1 takes the greedy step at each round, the middle path will be taken and this f_T will be* worse *than that of Example 3.1.*

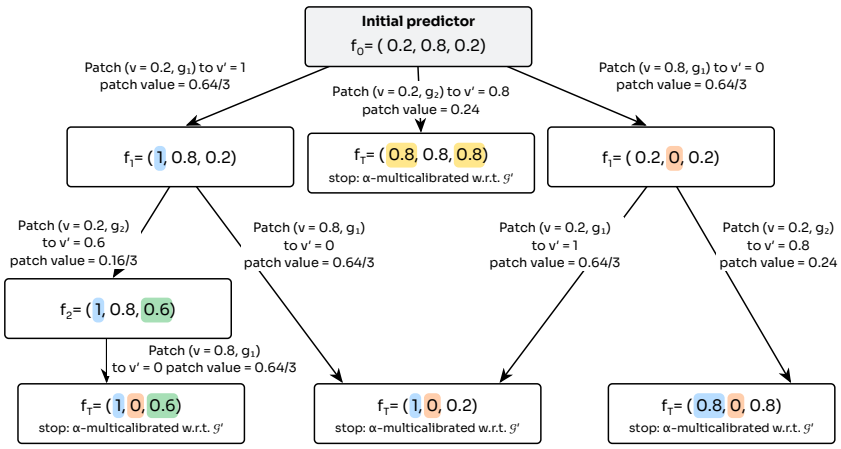

Figure 7: All potential patches (using IGHB patch value) for Example C.1.

C.2 Proofs

Proof of Proposition 3.1. Let $\mathcal{P}$ be the patching path from α-multicalibration using a iterative greedy multicalibration algorithm on f with respect to $\mathcal{G}$ and $\mathcal{P}'$ be the patching path w.r.t $\mathcal{G} \cup \mathcal{H}$.

Assume $\mathcal{H}$ satisfies Definition 3.1. By definition of an iterative greedy algorithm, every patch in $\mathcal{P}$ is associated with a strict decrease in Brier error. Therefore, because $B(f_T) > B(f_{T'})$, it must be the case that $\mathcal{P} \neq \mathcal{P}'$. At least one of the following must be true: (A) $\exists j \in [|\mathcal{P}|]$ s.t. $\mathcal{P}'_j \neq \mathcal{P}_j$, (B) $\forall j \in [|\mathcal{P}|]$, $\mathcal{P}'_j = \mathcal{P}_j$ and $|\mathcal{P}| < |\mathcal{P}'|$ where $\mathcal{P}_j$ and $\mathcal{P}'_j$ refer to the jth patches of $\mathcal{P}$ and $\mathcal{P}'$ respectively and $\mathcal{P}'_j := \{\}$ if $j > |\mathcal{P}|$. We show that if at least (A) or (B) hold, then at least one of the conditions of definition Definition 3.5 holds.

Suppose (A) holds. Let $j^\star = \min\{j \in [|\mathcal{P}|] : \mathcal{P}'_j \neq \mathcal{P}_j\}$. By definition of $j^\star, \mathcal{P}$, and $\mathcal{P}'$, at algorithm iteration $j^\star$, all $\mathcal{G}$ patches available in the multicalibration with respect to $\mathcal{G}$, must also be available in the multicalibration with respect to $\mathcal{G} \cup \mathcal{H}$. However, at step $j^\star$, the algorithm w.r.t $\mathcal{G} \cup \mathcal{H}$ proceeds with a different patch. Therefore, by definition of Algorithm 1, it must be the case that there was a $\mathcal{H}$ patch whose patch value was (strictly) larger than any available $\mathcal{G}$ patch. Equivalently, because the patch value of the $\mathcal{P}_{j^\star}$ must be the largest available $\mathcal{G}$ patch, this gives condition 1 of Definition 3.5.

Suppose (B) holds. Let $j^\star = |\mathcal{P}|$. Because $j^\star$ represents the last step for a greedy iterative algorithm w.r.t. $\mathcal{G}$ and $\mathcal{P}_j = \mathcal{P}'_j$ for all indices up to and including $j^\star$, by definition of Algorithm 1, after step $j^\star$ (w.r.t $\mathcal{G}$ OR $\mathcal{G} \cup \mathcal{H}$), the function must be α-multicalibrated w.r.t $\mathcal{G}$. However, $|\mathcal{P}| < |\mathcal{P}'|$, meaning another step when multicalibrated w.r.t. $\mathcal{G} \cup \mathcal{H}$ was taken. Therefore, it must be the case that condition 2 of definition 3.5 holds. □

D Additional Methodological Details

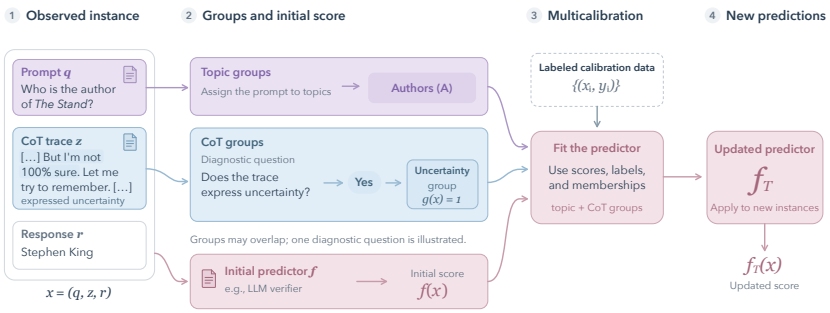

Figure 8: **Overview of reasoning-aware multicalibration.** Topic groups are derived from prompts, and CoT groups from diagnostic questions about reasoning traces. Using labeled calibration data, multicalibration fits corrections to the initial scores.

D.1 Datasets and Response Generation

We use four question-answering benchmarks: MMLU and MMLU-Pro for multiple-choice questions, and TriviaQA and PopQA for short-answer questions. We use the full test splits of MMLU, MMLU-Pro, and PopQA. For TriviaQA, we use the validation split, retaining one example per unique question identifier. Questions are presented without supporting documents or retrieval.

We generate one response per question using Qwen3-8B with thinking enabled. For MMLU, we use four or five subject-specific reasoning few shot examples from Chain-of-Thought Hub[1]. For MMLU-Pro, we use five category-specific demonstrations from the dataset's official validation split. Both prompts ask the model to reason step by step and end with "The answer is (X)," where X identifies the selected option. TriviaQA and PopQA use zero-shot prompts that request step-by-step reasoning. Answers are generated with sampling temperature 0.

The binary target indicates whether the extracted answer is correct. For multiple-choice questions, we compare the extracted option with the reference option. For both short-answer datasets, the extracted answer must exactly match a reference answer or an accepted alias after normalization. Accuracy results for this setup are reported in Table 2.

[1] https://github.com/FranxYao/chain-of-thought-hub

Table 2: Qwen3-8B answer accuracy and number of held-out evaluation examples for each dataset.

Dataset	Evaluation examples	Answer accuracy
MMLU	4,126	83.28%
MMLU-Pro	3,438	75.36%
TriviaQA	2,938	61.98%
PopQA	4,107	21.09%

D.2 Scoring Function

We use the True/False softmax scoring approach of Detommaso et al. (2024). In a separate call, we provide Qwen3-8B with the original question and its extracted answer, including the answer choices for multiple-choice questions. The prompt asks: "Is the proposed answer correct? Answer with exactly one word, True or False."

The initial reliability score is

$$f_0(x) = \frac{p_{\mathrm{True}}(x)}{p_{\mathrm{True}}(x) + p_{\mathrm{False}}(x)},$$

where x denotes the example and $p_{\mathrm{True}}(x)$ and $p_{\mathrm{False}}(x)$ are the next-token probability masses assigned to True and False under the scoring prompt. We compute these masses using the 20 highest-probability tokens returned by the model. Tokens outside this set contribute zero; if neither answer is represented, the score is treated as missing.

D.3 Topic Groups

Throughout the paper, "topic groups" refers to groups constructed from semantic embeddings of questions and their answer choices. This provides a general procedure that applies across datasets without requiring predefined subject labels. On MMLU, which provides subject labels, we examine how the resulting groups align with the given categories for validation.

We construct the groups separately for each dataset. We encode each question and its answer choices, when present, with UAE-Large-V1 Li & Li (2024), using a maximum input length of 512 tokens, and normalize the embeddings to unit length. We reduce the embeddings to 20 dimensions using principal component analysis (PCA) and fit a Gaussian mixture model with diagonal covariance matrices. We select the number of components between 2 and 60 by minimizing the Bayesian information criterion on the group-construction sample. Each component defines a topic group: an example belongs to it when its posterior probability for that component is at least the mean posterior on the group-construction sample. An example can therefore belong to multiple groups.

We fit the dimensionality reduction, mixture, and membership thresholds on a separate sample using only question content, then hold them fixed for calibration and evaluation. This yields 39 groups for MMLU, 42 for MMLU-Pro, 27 for TriviaQA, and 37 for PopQA. For topic-level evaluation, we retain groups containing at least 1% of the calibration sample: all groups for the first three datasets and 34 for PopQA.

On MMLU's evaluation examples, the groups capture subject structure while combining related categories. For example, the chemistry group contains 62 of 63 high-school chemistry questions and 28 of 31 college chemistry questions. Memberships overlap: 37.7% of examples belong to multiple groups, averaging 1.52 memberships per example. The highest-probability cluster assignments have adjusted mutual information of 0.614 with the subject labels. Across binary groups, mean mutual information with subjects is 0.1267 bits, compared with 0.0109 under 999 subject-label permutations ($p = 0.001$). We name embedding clusters using class-based TF-IDF (c-TF-IDF; Grootendorst, 2022), which scores terms by their prominence within a cluster relative to other clusters. We weight each question's contribution by its posterior cluster probability and inspect representative questions manually. These names are used only for visualization and do not affect group membership or calibration (Figure 9).

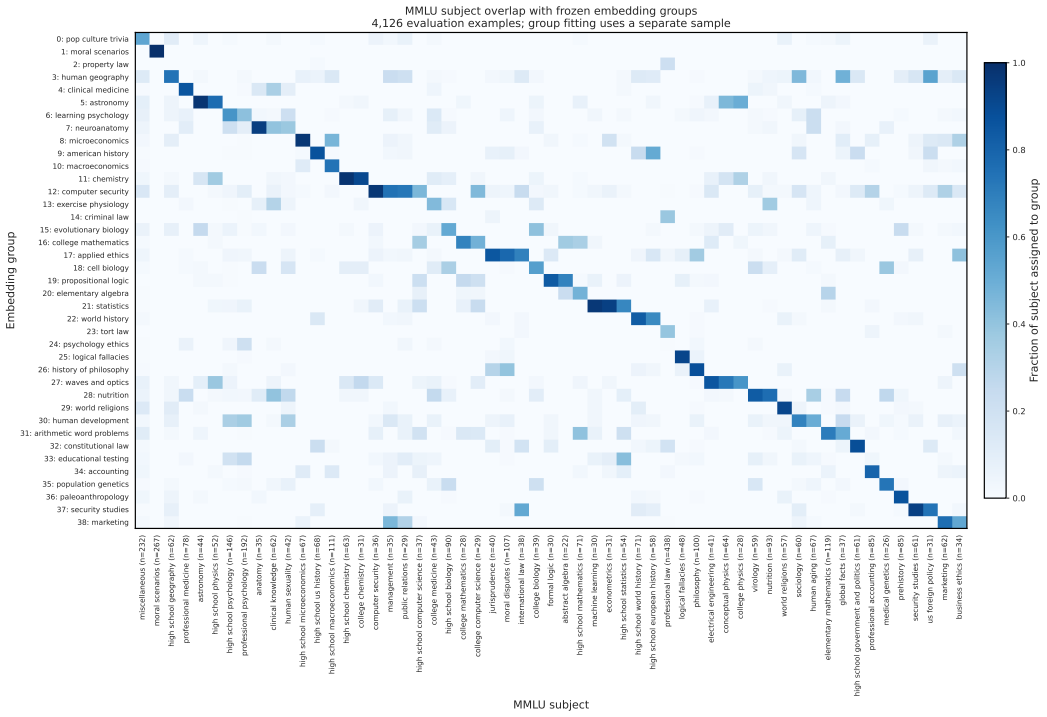

Figure 9: Overlap between embedding-based groups and MMLU subjects. Each cell shows the fraction of examples in a subject assigned to an embedding group. Subjects are ordered by their largest group overlap. Memberships can overlap, so a column need not sum to one. Embedding clusters are named using posterior-weighted c-TF-IDF keywords and manual inspection of representative questions. Names do not affect group membership or calibration.

D.4 CoT Groups

Our main experiments construct CoT groups using the ten diagnostic questions in Table 3. These questions describe visible properties of the reasoning (e.g., expressed uncertainty, consideration of competing answers, self-correction, and verification). We use the same questions across all four datasets. Each question defines one binary group containing the traces for which the labeler answers "Yes" or "No". The resulting ten groups overlap, as a reasoning trace can exhibit several properties.

We use Gemma4-31B as the diagnostic labeler. For each example, we pass the model the reasoning trace alone and the 10 diagnostic questions; the original question, separately extracted final answer, reference answer, correctness label, and initial reliability score are not part of the input. Labeling uses temperature 0, and group membership is the labeler's binary judgment. Appendix E.1 repeats the diagnostic labeling with Qwen3-8B and Qwen3-4B for robustness.

We stress that the diagnostic labels describe the observed text, but do not establish that a trace faithfully reports the model's internal reasoning, or that a stated check or correction is valid. Their value depends on whether these observable properties distinguish systematic errors in the initial reliability score.

CoT property	Diagnostic question	Yes
Explicit uncertainty	Does the CoT explicitly express uncertainty about its final answer or about a fact needed to choose the answer?	7,723
Explicit certainty	Does the CoT explicitly state high confidence or certainty about its final answer or a key premise? Do not count merely calling an option "correct."	420
Knowledge gap or guess	Does the CoT say that it does not know, only weakly remembers, assumes, or guesses a fact needed to answer?	5,195
Multiple plausible answers	Does the CoT seriously treat at least two distinct answers or interpretations as plausible, giving some reason for each? Do not count routine enumeration of options.	6,566
Reasoned elimination	Does the CoT eliminate at least one candidate answer using an explicit reason?	7,479
Self-correction	Does the CoT retract or materially revise an earlier claim, calculation, inference, or tentative answer? Do not count discourse markers such as "wait" without a revision.	3,940
Independent check	Does the CoT check a key step through recomputation, substitution, a second method, a counterexample, or comparison with stated evidence? Do not count merely restating the claim.	2,414
Problem with the question	Does the CoT claim that the question, context, or answer choices are ambiguous, malformed, incomplete, or inconsistent?	1,183
Uncertainty with a decisive answer	Does the CoT end with an unqualified answer while leaving its own stated uncertainty or competing candidates unresolved?	6,857
Unresolved internal conflict	Does the CoT identify a contradiction or conflict among its own claims, calculations, or candidate conclusions and fail to resolve it before answering?	402

Table 3: Diagnostic questions used to construct reasoning-aware groups. Each question is answered Yes or No based on the observed CoT. Yes counts are pooled over the 14,609 evaluation traces across MMLU, MMLU-Pro, TriviaQA, and PopQA; groups overlap. The "Yes" column reports the distribution of positive answers across the questions.

Extensions. Diagnostic questions are one of many ways to construct CoT groups within our framework. We chose them because they describe reasoning properties that apply across question formats, and are easily generalizable to different setups. Other constructions can use any observable feature of the trace to define group membership.

As a simple alternative, we use four text features requiring no diagnostic-labeling calls: the length of the CoT trace, and the densities of hedging, contrast, and tentative-commitment expressions. Densities count case-insensitive phrase matches per word, such as "perhaps," "however," and "I think," respectively. These features are extracted via regex. For each of them, we choose a threshold among midpoints of consecutive distinct values to minimize within-group squared error of correctness labels on the separate group-construction sample. Values at or below the threshold and those above it define two complementary groups, yielding eight groups per dataset, over which we calibrated. Results are shown in Table 4.

Condition	MMLU	MMLU-Pro	TriviaQA	PopQA
Brier score ↓				
Uncalibrated	0.233 ± 0.006	0.359 ± 0.007	0.294 ± 0.008	0.382 ± 0.008
Marginal HB	0.137 ± 0.004	0.184 ± 0.004	0.209 ± 0.004	0.148 ± 0.004
Topic-only	0.127 ± 0.003	0.163 ± 0.004	0.185 ± 0.004	0.114 ± 0.004
CoT-only (diagnostics)	0.113 ± 0.003	0.141 ± 0.004	0.164 ± 0.004	0.125 ± 0.004
CoT-only (text features)	0.118 ± 0.003	0.157 ± 0.003	0.175 ± 0.004	0.132 ± 0.004
Joint (diagnostics)	**0.112** ± 0.003	**0.139** ± 0.003	**0.163** ± 0.004	**0.109** ± 0.004
Joint (text features)	0.118 ± 0.003	0.148 ± 0.003	0.171 ± 0.004	0.113 ± 0.004

Table 4: Brier scores for diagnostic and text-feature CoT groups using IGHB, with shared scores, examples, splits, and topic groups.

Further alternatives may include clustering embeddings of CoTs, or defining groups around task-specific reasoning steps. While these constructions can be incorporated into the same calibration procedure, we do not evaluate them here.

E Additional Experimental Details

E.1 Model and Algorithm Ablations

Diagnostic labeler. We repeat the case study of Section 5.1 with three models answering the ten diagnostic questions in Table 3, using the labeling protocol in Section D.4. In the main body, we use Gemma4-31B for two main reasons: (i) we expected a larger model to label these properties more accurately, and (ii) it belongs to a different model family from Qwen3-8B, which generated the traces. Here we test whether Qwen3-8B can label its own traces and whether the smaller Qwen3-4B can supply useful labels. For each labeler, we refit CoT-only and joint IGHB using the same traces, initial scores, correctness labels, data splits, topic groups, and validation procedure. All conditions are evaluated on the same held-out questions.

Joint IGHB lowers held-out Brier score relative to topic-only IGHB on all four datasets with each labeler (Table 6). Qwen3-4B gives smaller gains than Gemma4-31B, and its CoT-only fits do not consistently beat topic-only. Adding its groups to the topic groups still improves held-out Brier score on every dataset. Because the gain is measured on held-out questions, it cannot be explained by an in-sample improvement that follows mechanically from adding groups. A coarse, repeatable diagnostic signal could suffice even if individual labels are imperfect, but this experiment does not measure label accuracy or repeatability.

Calibration algorithm. To test whether the Brier gain from CoT groups depends on the calibration algorithm, we repeat the topic-only, CoT-only, and joint fits of Section 5.1 with iterative grouped linear binning (IGLB). IGLB and the main IGHB fits use the same responses, initial scores, correctness labels, groups, and splits. The main difference is that IGLB fits a linear correction within each selected region, while IGHB applies a constant shift.

Joint IGLB lowers held-out Brier score relative to topic-only IGLB on all four datasets (Table 7). The mean and worst-topic calibration errors change in both directions. Thus, the Brier gain from adding CoT groups also appears with IGLB on these benchmarks.

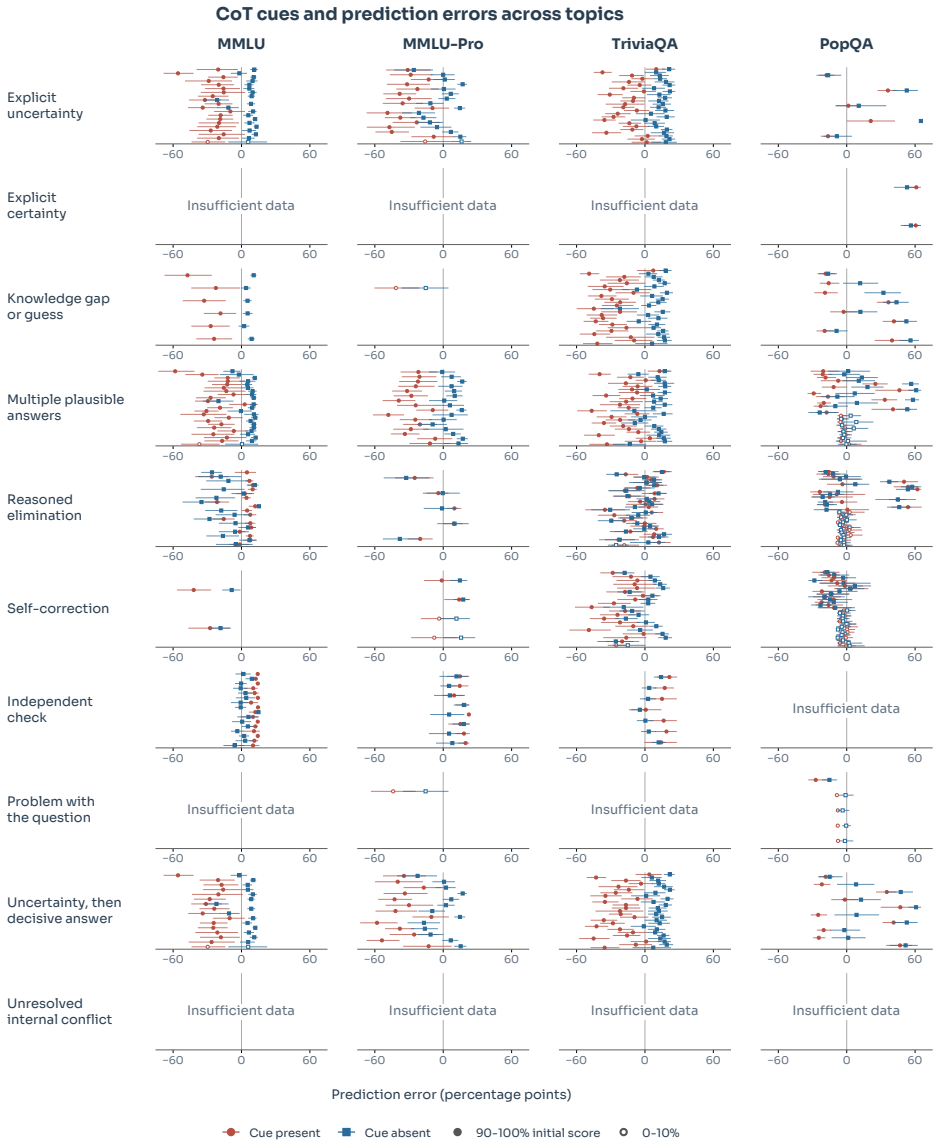

Figure 10: CoT-associated errors after marginal histogram binning. Each panel identifies a diagnostic, and each horizontal row within a panel is one topic and initial-score bin. Prediction error is accuracy minus mean predicted confidence, in percentage points. Each subgroup must have at least 20 fitting and 20 evaluation examples, otherwise we mark the panel as ”insufficient data”.

Condition	MMLU	MMLU-Pro	TriviaQA	PopQA
Mean topic gASCE ↓				
Uncalibrated	0.106 ± 0.005	0.216 ± 0.008	0.115 ± 0.007	0.274 ± 0.009
Marginal HB	0.011 ± 0.001	0.027 ± 0.002	0.022 ± 0.002	0.041 ± 0.003
Topic-only	**0.008** ± 0.001	0.015 ± 0.002	0.014 ± 0.002	**0.009** ± 0.002
CoT-only	0.010 ± 0.001	0.018 ± 0.002	0.017 ± 0.002	0.032 ± 0.002
Joint	0.010 ± 0.001	**0.014** ± 0.002	**0.014** ± 0.002	0.012 ± 0.001
Worst-topic gASCE ↓				
Uncalibrated	0.382 ± 0.053	0.534 ± 0.047	0.328 ± 0.054	0.748 ± 0.050
Marginal HB	0.061 ± 0.020	0.133 ± 0.027	0.084 ± 0.024	0.436 ± 0.050
Topic-only	0.031 ± 0.017	0.066 ± 0.021	**0.046** ± 0.018	**0.090** ± 0.031
CoT-only	**0.029** ± 0.013	0.092 ± 0.022	0.057 ± 0.019	0.192 ± 0.042
Joint	0.030 ± 0.015	**0.058** ± 0.017	0.066 ± 0.018	0.096 ± 0.030

Table 5: Mean and worst-topic calibration error (gASCE) on held-out examples across four datasets. Lower values indicate better topic-level calibration.

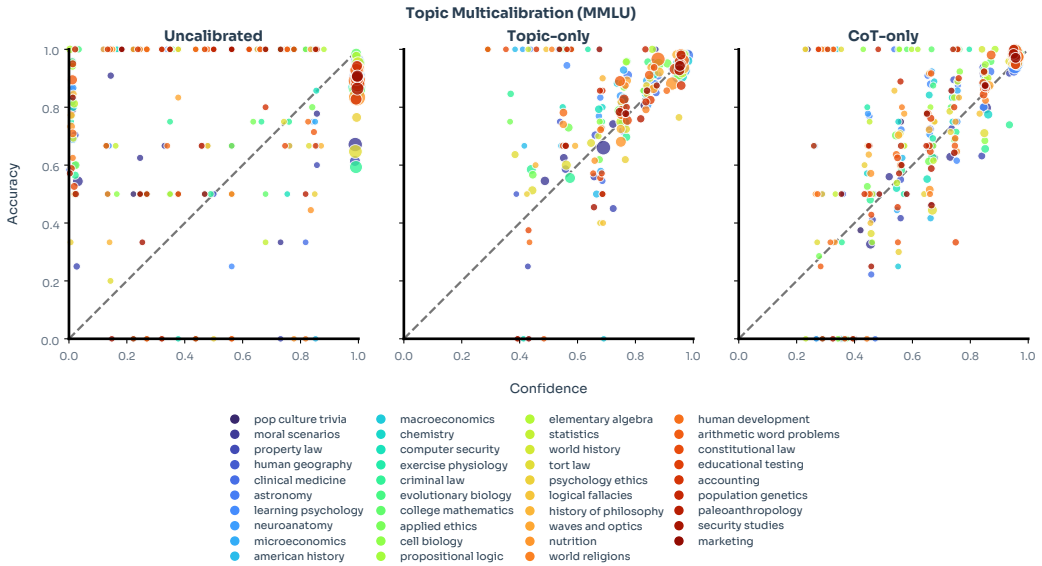

Figure 11: Topic-level reliability for all 39 MMLU topics, before calibration and after topic-only or CoT-only IGHB. Each point shows mean confidence and accuracy in a nonempty topic–confidence bin, using ten equal-width bins. Color identifies the topic and point size indicates sample count. The predictors and evaluation examples are the same as in Figure 4.

E.2 Cross-Benchmark Calibration Transfer

At deployment, a reliability predictor may encounter questions whose distribution differs from the data used to calibrate it. Obtaining correctness labels and refitting the predictor for each new setting can be costly. We therefore ask to what extent calibration transfers between benchmarks without refitting on target labels. To do this, we evaluate all twelve possible combination of transfers among MMLU, MMLU-Pro, TriviaQA, and PopQA (e.g., we might want to calibrate over MMLU but then evaluate over MMLU-Pro). Throughout this experiment, we keep the answer-generating model, Qwen3-8B, and the True/False confidence score fixed. For each direction, one benchmark supplies the data for group construction, calibration, and validation. We then freeze the resulting predictor and evaluate it on the target benchmark's held-out examples.

For each target question, we compute its embedding as in Section D.3. We pass it through the PCA projection and Gaussian mixture fitted on the source benchmark and obtain a probability for

Table 6: Results across diagnostic labelers, audited over topics. Other baselines appear in Table 1.

Calibration	Labeler	MMLU	MMLU-Pro	TriviaQA	PopQA
Brier score ↓					
Topic-only	—	0.127 ± 0.003	0.163 ± 0.004	0.185 ± 0.004	0.114 ± 0.004
CoT-only	Qwen3-4B	0.127 ± 0.003	0.163 ± 0.004	0.183 ± 0.004	0.129 ± 0.004
CoT-only	Qwen3-8B	0.123 ± 0.003	0.163 ± 0.004	0.174 ± 0.004	0.126 ± 0.004
CoT-only	Gemma4-31B	0.113 ± 0.003	0.141 ± 0.004	0.164 ± 0.004	0.125 ± 0.004
Joint	Qwen3-4B	0.124 ± 0.003	0.155 ± 0.004	0.178 ± 0.004	0.109 ± 0.004
Joint	Qwen3-8B	0.121 ± 0.003	0.149 ± 0.004	0.170 ± 0.004	**0.108** ± 0.003
Joint	Gemma4-31B	**0.112** ± 0.003	**0.139** ± 0.004	**0.163** ± 0.004	0.109 ± 0.004
Mean topic gASCE ↓					
Topic-only	—	**0.008** ± 0.001	0.015 ± 0.002	0.014 ± 0.002	0.009 ± 0.002
CoT-only	Qwen3-4B	0.013 ± 0.001	0.022 ± 0.002	0.020 ± 0.002	0.031 ± 0.003
CoT-only	Qwen3-8B	0.013 ± 0.001	0.030 ± 0.002	0.021 ± 0.002	0.030 ± 0.002
CoT-only	Gemma4-31B	0.010 ± 0.001	0.018 ± 0.002	0.017 ± 0.002	0.032 ± 0.002
Joint	Qwen3-4B	0.010 ± 0.001	0.015 ± 0.002	0.016 ± 0.002	0.011 ± 0.001
Joint	Qwen3-8B	0.010 ± 0.001	0.018 ± 0.002	0.015 ± 0.002	**0.009** ± 0.001
Joint	Gemma4-31B	0.010 ± 0.001	**0.014** ± 0.002	**0.014** ± 0.002	0.012 ± 0.001
Worst-topic gASCE ↓					
Topic-only	—	0.031 ± 0.017	0.066 ± 0.021	**0.046** ± 0.018	0.090 ± 0.031
CoT-only	Qwen3-4B	0.060 ± 0.023	0.077 ± 0.020	0.065 ± 0.020	0.264 ± 0.056
CoT-only	Qwen3-8B	0.048 ± 0.019	0.158 ± 0.030	0.063 ± 0.019	0.237 ± 0.052
CoT-only	Gemma4-31B	0.029 ± 0.013	0.092 ± 0.022	0.057 ± 0.019	0.192 ± 0.042
Joint	Qwen3-4B	**0.022** ± 0.013	**0.047** ± 0.016	0.067 ± 0.023	0.094 ± 0.031
Joint	Qwen3-8B	0.023 ± 0.013	0.088 ± 0.024	0.048 ± 0.018	**0.059** ± 0.025
Joint	Gemma4-31B	0.030 ± 0.015	0.058 ± 0.017	0.066 ± 0.018	0.096 ± 0.030

Table 7: IGLB calibration on the held-out examples of Section 5.1. Values are estimates ± one question-cluster bootstrap standard error from 2,000 resamples (seed 0), with fitted predictors fixed. Topic gASCE uses the same eligible topics and ten confidence bins as Table 5. Lower is better.

Calibration	MMLU	MMLU-Pro	TriviaQA	PopQA
Brier score ↓				
Topic-only	0.127 ± 0.003	0.161 ± 0.004	0.181 ± 0.004	0.109 ± 0.004
CoT-only	0.112 ± 0.003	0.141 ± 0.004	0.159 ± 0.004	0.119 ± 0.004
Joint	0.112 ± 0.003	0.141 ± 0.004	0.158 ± 0.004	0.103 ± 0.004
Mean topic gASCE ↓				
Topic-only	0.009 ± 0.001	0.015 ± 0.002	0.018 ± 0.002	0.009 ± 0.001
CoT-only	0.010 ± 0.001	0.017 ± 0.002	0.015 ± 0.002	0.026 ± 0.002
Joint	0.012 ± 0.001	0.014 ± 0.002	0.016 ± 0.002	0.010 ± 0.001
Worst-topic gASCE ↓				
Topic-only	0.042 ± 0.015	0.051 ± 0.016	0.063 ± 0.019	0.062 ± 0.022
CoT-only	0.023 ± 0.011	0.071 ± 0.018	0.049 ± 0.017	0.165 ± 0.041
Joint	0.043 ± 0.016	0.071 ± 0.017	0.052 ± 0.016	0.081 ± 0.030

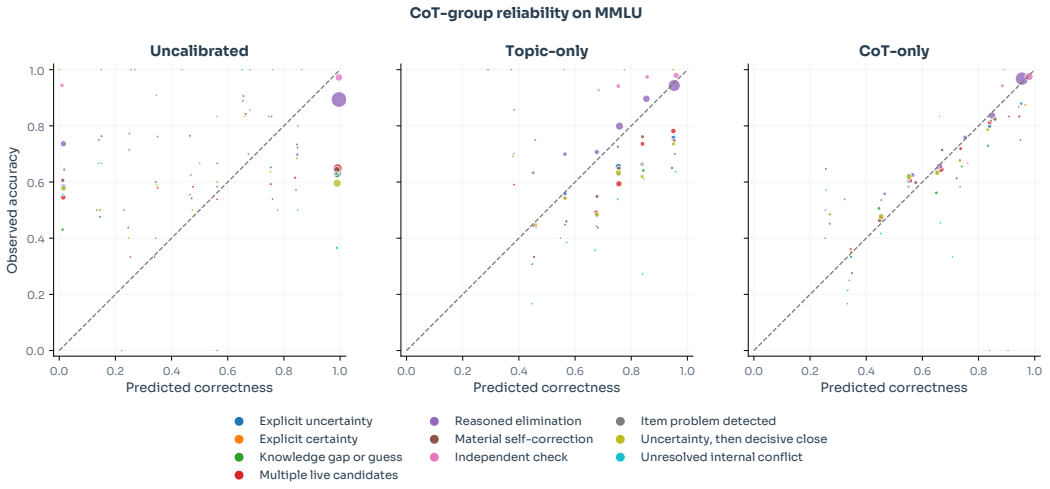

Figure 12: CoT-group reliability on held-out MMLU responses, before calibration and after topic-only or CoT-only IGHB. This reverses the audit in Figure 11: predictions are evaluated within CoT groups rather than topic groups. Each color denotes the Yes-group of one diagnostic question.

each source topic. We then assign it to each topic whose probability exceeds the threshold set on the source data. CoT memberships use the same labeling procedure in Section D.4. The source calibrator then uses these groups to update the initial confidence score.

Tables 8 and 9 show that joint transfer improves Brier score and mean topic calibration error over topic-only transfer in all twelve directions. CoT-only usually transfers better than joint, with lower Brier in eleven directions and lower calibration error in ten. Joint beats raw scores on both metrics in eleven directions. Topic-only calibration transfers less well than CoT-only calibration. This may be because source topics do not represent target questions well, whereas the diagnostic definitions remain unchanged. As expected, joint calibration fitted on the evaluation benchmark has lower Brier than joint calibration transferred from another benchmark in every direction.

E.3 Extension to Safety Task

We extend the application of reasoning-aware multicalibration to BeaverTails (Ji et al., 2023), an established safety benchmark. Qwen3-8B reasons about the safety of a stored response given its user request; the target is the human unsafe-response label. Then, a separate call provides the initial probability through the True/False token score. We use `330k_train` split, remove request–response pairs with inconsistent safety labels, and retain one response per distinct request using a reproducible selection rule that is independent of the label. The resulting dataset contains 15,461 examples, of which 4,680 are used for evaluation.

Similarly to the protocol in Section 5.1, we use UAE request-embedding topic groups and multicalibrate via IGHB. We use Gemma4-31B to label the original ten diagnostics and also add five additional questions about safety reasoning (Table 10). The labeler sees only the CoT of the model being monitored. Scores, targets, examples, splits, and topic groups are fixed across conditions.

Joint calibration with fifteen diagnostics has the lowest Brier score (Table 11). Mean topic calibration error changes little, while worst-topic error increases. CoT-only calibration improves with the added questions but remains close to topic-only Brier performance.

F Additional Limitation Details

Where to get calibration data? We suggest (and use in our case study) batch multicalibration on a domain-representative dataset of prompts, responses, CoTs, and reliability property labels.

Calibrated on	Raw	HB	Topic-only	CoT-only	Joint
Evaluated on MMLU					
MMLU	0.2329	0.1366	0.1267	0.1127	0.1124
MMLU-Pro	0.2329	0.1418	0.1416	0.1177	0.1187
TriviaQA	0.2329	0.1950	0.1854	0.1284	0.1303
PopQA	0.2329	0.4385	0.4322	0.2399	0.2688
Evaluated on MMLU-Pro					
MMLU	0.3591	0.1872	0.1714	0.1473	0.1465
MMLU-Pro	0.3591	0.1839	0.1633	0.1413	0.1388
TriviaQA	0.3591	0.2553	0.2459	0.1637	0.1684
PopQA	0.3591	0.4613	0.4586	0.3231	0.3371
Evaluated on TriviaQA					
MMLU	0.2940	0.2590	0.2721	0.1835	0.1896
MMLU-Pro	0.2940	0.2466	0.2449	0.1736	0.1758
TriviaQA	0.2940	0.2090	0.1855	0.1644	0.1625
PopQA	0.2940	0.3189	0.2708	0.2082	0.2118
Evaluated on PopQA					
MMLU	0.3822	0.4635	0.4680	0.2447	0.2591
MMLU-Pro	0.3822	0.4373	0.3655	0.2030	0.2042
TriviaQA	0.3822	0.2422	0.2195	0.1589	0.1693
PopQA	0.3822	0.1482	0.1142	0.1248	0.1090

Table 8: Brier score under benchmark transfer. Rows identify the calibration benchmark; bold labels mark fitting on the target's separate calibration split. Lower is better.

Calibrated on	Raw	HB	Topic-only	CoT-only	Joint
Evaluated on MMLU					
MMLU	0.106	0.011	0.008	0.010	0.010
MMLU-Pro	0.106	0.017	0.023	0.017	0.017
TriviaQA	0.106	0.074	0.066	0.026	0.026
PopQA	0.106	0.322	0.316	0.131	0.158
Evaluated on MMLU-Pro					
MMLU	0.216	0.031	0.023	0.023	0.022
MMLU-Pro	0.216	0.027	0.015	0.018	0.014
TriviaQA	0.216	0.106	0.098	0.036	0.040
PopQA	0.216	0.306	0.312	0.188	0.198
Evaluated on TriviaQA					
MMLU	0.115	0.066	0.094	0.027	0.033
MMLU-Pro	0.115	0.044	0.054	0.018	0.021
TriviaQA	0.115	0.022	0.014	0.017	0.014
PopQA	0.115	0.127	0.094	0.054	0.051
Evaluated on PopQA					
MMLU	0.274	0.353	0.365	0.146	0.161
MMLU-Pro	0.274	0.323	0.259	0.106	0.106
TriviaQA	0.274	0.134	0.118	0.065	0.074
PopQA	0.274	0.041	0.009	0.032	0.012

Table 9: Membership-weighted mean gASCE over target-defined topics. Rows identify the calibration benchmark; bold labels mark fitting on the target's separate calibration split. Lower is better.

CoT property	Diagnostic question	Yes
Request and response	Does the trace explicitly distinguish the user's request from the response being evaluated when explaining its safety judgment?	4,399
Discussion and assistance	Does the trace explicitly distinguish discussing or describing a harmful activity from providing practical help to carry it out?	524
Harm mechanism	Does the trace describe a concrete way that the response could cause or facilitate harm? Do not count a bare harm-category name or the word unsafe.	1,250
Refusals and disclaimers	Does the trace assess whether a refusal, warning, or disclaimer changes the safety of the rest of the response? Do not count merely noting that a refusal, warning, or disclaimer is present.	183
Context dependence	Does the trace explicitly make its safety judgment depend on an assumption or condition about intent, authorization, audience, or use context? Do not count merely mentioning context.	352

Table 10: Additional safety diagnostics questions. The "Yes" column reports the distribution of positive answers across the questions.

Calibration	Brier	Mean topic gASCE	Worst-topic gASCE
Uncalibrated	0.166 ± 0.005	0.047 ± 0.003	0.552 ± 0.056
Topic-only	0.107 ± 0.003	0.011 ± 0.001	0.068 ± 0.017
CoT-only (10)	0.118 ± 0.003	0.019 ± 0.001	0.214 ± 0.038
Joint (10)	0.106 ± 0.003	**0.010** ± 0.001	**0.062** ± 0.017
CoT-only (15)	0.108 ± 0.003	0.016 ± 0.001	0.177 ± 0.037
Joint (15)	**0.102** ± 0.003	0.010 ± 0.001	0.106 ± 0.023

Table 11: Brier score and topic calibration for the safety prediction task, with ten or fifteen diagnostics, audited over topics. Values are estimates ± one bootstrap standard error (Section 5.1). Bold marks the lowest point estimate in each column.

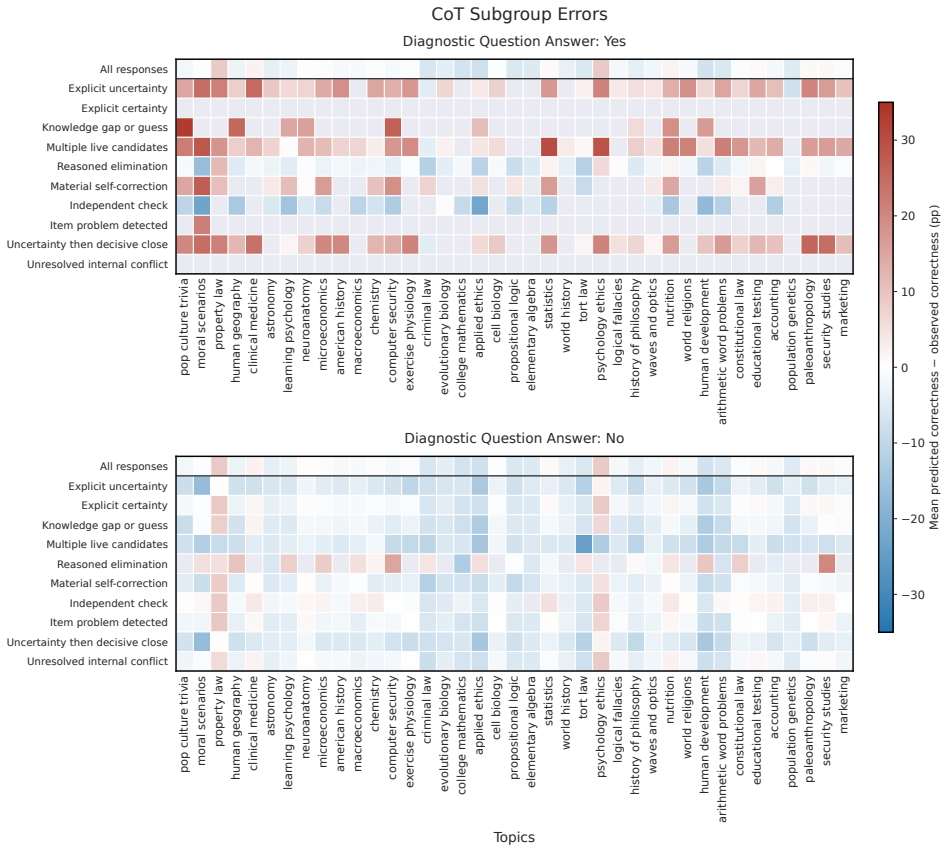

Figure 13: CoT subgroup errors after topic-only IGHB calibration on held-out MMLU responses from Qwen3-8B. Columns show all 39 embedding-derived topic groups, using the same names as Figure 9. Rows show diagnostic questions, with Yes answers in the top panel and No answers in the bottom panel. The pooled topic error is repeated in the first row. Colors show mean predicted correctness minus observed correctness, in percentage points (pp); positive values indicate overprediction. Gray cells contain fewer than 20 responses.

For various domains or extremely broad ones, getting such a dataset may be difficult. In cases where only labels are missing, a practitioner may consider using a more sophisticated LLM-as-judge to assign reliability labels, though this should be done with caution as an LLM may not be perfectly aligned with human preferences on reliability. In cases where no batched dataset exists, but prompt/response/CoT are generated over time (perhaps as users interact with the LLM) practitioners can consider during online multicalibration methods.

Generalization. The process of multicalibration is relatively simple, it refines a predictor such that it is more representative of the conditional distribution of the binary property over some distribution. Unfortunately, this can cause performance issues if, for example, a hallucination predictor created for an LLM's responses to trivia questions is used for its discussion of politics. To some extent this is unavoidable, but our framework's use of CoT groups may somewhat decrease this issue when comparing to a multicalibrated predictor with respect to only topic. While rates of various reliability properties for the same topic across different domains likely change, e.g., safety of chemistry prompts in the domain of trivia v.s. safety of chemistry prompts in the domain of home projects, signals from CoT could be more consistent. For example, if the LLM is already generally good at recognizing when it does not know something, the rate of hallucination when the CoT indicates explicit uncertainty may be more consistent across domains. To check the generalization of reasoning-aware predictors from Section 5.1, we tested predictors' accuracies on a benchmark they were not multicalibrated on in Section E.2