



# LLM validity via enhanced conformal prediction: conditional guarantees, score learning, and response-level calibration under dependent claims

Davide Beltrame<sup>1</sup>

<sup>1</sup>Bocconi Students for Machine Learning, Bocconi University, Milan, Italy  
[davide.beltrame@studbocconi.it](mailto:davide.beltrame@studbocconi.it)

February 7, 2026

## Abstract

Finite-sample validity guarantees for large language model (LLM) outputs are attractive because they are post-hoc and model-agnostic, but they are fragile when prompts are heterogeneous and factuality signals are noisy. Cherian et al. [2024] propose enhanced conformal methods for *factual filtering* that (i) replace marginal guarantees with function-class conditional guarantees and (ii) improve utility via level-adaptive calibration and conditional boosting that differentiates through conditional conformal cutoffs. This discussion paper reviews the conformal prediction and LLM factuality context, presents the selected paper and proposes a future direction: *response-level conformalization under dependent claims*. The key idea is to treat an entire response as the exchangeable unit, use blocked calibration that keeps all claims from a response together, and calibrate response-level tail losses conditionally on prompt/response features to align guarantees with user-facing risk when claim errors are dependent.

## Introduction

Large language models (LLMs) can produce fluent answers that contain factual errors. A user-facing system therefore needs a principled way to decide how much of a response to show, when to hedge, and when to abstain. Conformal prediction turns a post-hoc score into a finite-sample guarantee under exchangeability. In LLM factuality filtering, the guarantee is only meaningful relative to a claim extractor and a verifier, as marginal validity can hide subgroup failures across prompt types.

This work situates Cherian et al. [2024] in conditional and risk-control conformal theory, explains their conditional/level-adaptive/boosting contributions, and proposes dependence-aware response-level tail-risk targets.

## 1 Literature review

**Conformal prediction and conditional validity.** Conformal prediction provides distribution-free predictive guarantees under exchangeability [Vovk et al., 2005, Shafer and Vovk, 2008]. In a standard supervised setting with exchangeable data points  $(X_i, Y_i)$ , where  $X_i \in \mathcal{X}$  denotes the input features (e.g., prompts) and  $Y_i \in \mathcal{Y}$  denotes the target labels (e.g., ground truth), conformal methods construct a set-valued predictor  $\widehat{C}(\cdot)$  such that

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - \alpha, \quad (1)$$

for any sample size and without parametric assumptions [Shafer and Vovk, 2008, Lei et al., 2018]. The guarantee is finite-sample and does not require correctness of the underlying model, only exchangeability and symmetry of the procedure.

Split (inductive) conformal is the workhorse variant for modern ML [Papadopoulos et al., 2002, Lei et al., 2018]. It trains a predictor on one fold and calibrates a threshold on the other fold using a scalar *nonconformity* score. This separation makes conformal wrappers computationally feasible for large models, but introduces sensitivity to the calibration distribution and to score quality. Exact distribution-free *conditional coverage*,

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) \mid X_{n+1} = x) \geq 1 - \alpha \quad \text{for all } x, \quad (2)$$

is impossible in general without vacuous sets for continuous problems [Barber et al., 2019]. This negative result is central for LLM safety: it formalizes why guaranteed safety for every prompt is not attainable without strong modeling assumptions.

A tractable relaxation is to require validity *over a function class* or over a collection of groups. [Gibbs et al., 2024] develop conditional conformal procedures that provide guarantees of the form

$$\mathbb{E}\left[f(X_{n+1})\left(\mathbf{1}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} - (1 - \alpha)\right)\right] = 0, \quad \forall f \in \mathcal{F}. \quad (3)$$

[Cherian et al., 2024] work with a finite-dimensional linear class  $\mathcal{F} = \{\Phi(\cdot)^\top \beta : \beta \in \mathbb{R}^d\}$  and a randomized cutoff; under exchangeability and permutation-invariant (symmetric) optimization, the randomized rule yields exact moment equalities (otherwise one typically gets conservative inequalities due to ties). The message is that conditionality must be specified: the guarantee is only as rich as  $\mathcal{F}$ .

A related line of work uses “many-slices” constraints inspired by algorithmic fairness. Batch multi-valid conformal prediction enforces coverage across many overlapping groups in a batch setting [Jung et al., 2022]. This viewpoint is useful for LLMs because subgroup failures are often the core risk, but it also highlights a design burden: one must define, approximate, or learn a set of slices that is meaningful for deployment.

**Risk control.** Coverage is not always the most decision-relevant target. Risk-control frameworks replace “ $Y \in \widehat{C}(X)$ ” with control of a (typically bounded, monotone) loss. Learn-then-Test (LTT) frames calibration as hypothesis testing over candidate parameters [Angelopoulos et al., 2025a]. Conformal Risk Control (CRC) targets expected loss control for monotone losses [Angelopoulos et al., 2025b]. In LLM factuality, this matters because users may tolerate small errors in low-stakes contexts but demand strict behavior in high-stakes domains. Risk control makes this explicit by letting the user choose a loss functional (e.g., expected number of factual errors) rather than only a zero-error event. The selected paper sits between these views. It keeps the split-conformal “high-probability” framing (probability of satisfying a loss constraint), but extends it with conditional structure and with level adaptivity that trades off strictness and utility [Cherian et al., 2024].

We focus on *factual filtering*, where conformal calibration wraps a claim-level scoring pipeline to control a response-level loss [Mohri and Hashimoto, 2024, Cherian et al., 2024]. Other conformal interfaces for LLMs (set generation, RAG certification, and discrete QA conformalization) are summarized in Appendix B.

**Factuality pipelines.** Long-form factuality evaluation typically decomposes a response into atomic claims and labels each claim using a verifier (human annotation, LLM-as-judge, or retrieval-augmented verification) [Min et al., 2023, Wei et al., 2024]. These labels make conformal calibration possible, but they also fix the target of any guarantee: validity is always *relative to the chosen extraction and verifier*, not to an abstract notion of truth. This matters in high-stakes and heterogeneous domains (e.g., medical long-form QA), where label noise, evidence shift, and prompt heterogeneity are often the dominant failure modes. Appendix C lists concrete benchmark examples and common verifier/score choices.

**Learning scores for efficiency.** Conformal guarantees depend on the quality of the conformity score: weak scores force aggressive filtering and collapse utility. A recent theme is to learn scores that improve efficiency under conformal constraints [Stutz et al., 2021, Bai et al., 2022, Einbinder et al., 2022]. [Cherian et al., 2024] contribute a conditional analogue via *conditional boosting*, i.e., differentiating through a *conditional* conformal calibration step.

Output	Conformal Factuality Fixed level: 90%	Our Method Adaptive level: 63%
The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.	The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.	The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.

**Figure 1:** Illustrative example of conformal factuality filtering. The left panel displays the output of GPT-3.5-Turbo for the prompt “How often is ashingles vaccine required?” The first filtered output (center) is calibrated using the frequency score (see Appendix E.1 in [Cherian et al., 2024]) and the marginally valid conformal factuality method of [Mohri and Hashimoto, 2024] at a fixed level of 90%. The second filtered output (right) is calibrated using a score obtained via their conditional boosting procedure at a level of 63%, which is chosen and calibrated using their adaptive method to approximately ensure that at least 70% of the claims are retained. Both procedures provide distribution-free guarantees that are marginal over the calibration draw (and exact under randomized tie-breaking); the reported level should be interpreted as calibrated over the slices encoded by  $\mathcal{F}$ , not as a pointwise per-prompt probability.

Two design choices dominate the interpretation of conformal guarantees in LLM pipelines. First, one must specify *how* conditionality is enforced (groups, function classes, or multivalid constraints). Second, one must choose a loss and a calibration unit that match the user-facing object (a response, not a bag of claims). Section 2 builds the selected paper around these choices.

## 2 Article presentation

Cherian et al. [2024] study a factuality filtering pipeline for LLM outputs. Each data point is a prompt–response pair  $(P_i, R_i)$  together with an extracted claim list and claim labels. Formally, let

$$D_i = (P_i, R_i, C_i, W_i), \quad i = 1, \dots, n,$$

where  $C_i = \{C_{ij}\}_{j=1}^{k_i}$  is the list of claims extracted from  $R_i$ , and  $W_i = \{W_{ij}\}_{j=1}^{k_i}$  indicates which claims are factual under a verifier. In their experiments, claim extraction is done by a state-of-the-art commercial LLM (e.g., GPT-4o), while labels come from either an LLM-based verifier against a pseudo-ground-truth response (medical QA) or a retrieval-based pipeline inspired by FActScore (Wikipedia biographies) [Min et al., 2023].

A claim scoring function assigns a real-valued confidence to each claim. The selected paper uses the notation  $p(P_i, C_{ij})$  in the baseline and a parameterized score in the boosting section. To avoid inconsistent arguments, we write a generic score as  $s_\theta(P_i, R_i, c) \in \mathbb{R}$ , with the understanding that  $p(P_i, C_{ij})$  is a special case where the score ignores parts of  $(P_i, R_i)$ .

**Definition 1 (Filtering operator)** Fix a (possibly learned) claim score  $s_\theta$ . Given a prompt–response pair  $(P, R)$  with extracted claims  $C$  and a threshold  $\tau \in \mathbb{R}$ , define

$$F(C; \tau) = \{c \in C : s_\theta(P, R, c) \geq \tau\}. \quad (4)$$

When  $(P, R)$  is clear from context (e.g.,  $C = C(P, R)$ ), we write  $F(C; \tau)$  without explicitly displaying  $(P, R)$ .

A key interpretability point is that the evaluation target is response-level even though it is defined via claims. Users consume an entire answer, and the guarantee concerns a global event such as “no false retained claim” or “at most  $k$  false retained claims”. The motivating example in Cherian et al. [2024] compares fixed-level filtering to an adaptive-level, boosted-score filtering for a medical prompt (Fig. 1).

## 2.1 Baseline: Mohri–Hashimoto conformal factuality filtering

The closest baseline is conformal factuality filtering from [Mohri and Hashimoto \[2024\]](#), which calibrates a threshold so that the filtered output contains no false claims with high probability. The target can be written as follows.

**Definition 2 (Conformal factuality goal)** *Let  $\hat{\tau}$  be a data-dependent threshold calibrated on a held-out set, and define the filtered output  $\hat{F}(C_{n+1}) := F(C_{n+1}; \hat{\tau})$ . The goal is*

$$\mathbb{P}\left(\exists c \in \hat{F}(C_{n+1}) \text{ s.t. } W_{n+1}(c) = 0\right) \leq \alpha. \quad (5)$$

Whenever convenient, we identify  $W_i = \{W_{ij}\}_{j=1}^{k_i}$  with the map  $W_i : C_i \rightarrow \{0, 1\}$  defined by  $W_i(C_{ij}) := W_{ij}$ . Mohri–Hashimoto define a response-level conformity score capturing the minimum threshold that would remove all false retained claims:

**Definition 3 (Response-level conformity score for “no false retained”)**

$$S(C_i, W_i) = \inf \{\tau \in \mathbb{R} : \forall c \in F(C_i; \tau), W_i(c) = 1\}. \quad (6)$$

Split conformal then sets  $\hat{\tau}$  to a suitable upper quantile of  $S(C_i, W_i)_{i=1}^n$ . Under exchangeability of  $(P_i, R_i, C_i, W_i)$ , this yields marginal validity for the event in (5) [[Mohri and Hashimoto, 2024](#), [Vovk et al., 2005](#), [Shafer and Vovk, 2008](#)].

[Cherian et al. \[2024\]](#) use this baseline as the main reference point. They emphasize two problems. First, the guarantee is marginal and can hide subgroup failures across prompt characteristics. Second, if the score is imperfect, enforcing a strict event like “no false retained” can remove many true claims, collapsing utility.

If claim scores can tie, the infimum in (6) need not be attained under the non-strict rule  $F(C_i; \tau) = \{c \in C_i : s_\theta(P_i, R_i, c) \geq \tau\}$ . Throughout, either assume the claim scores are a.s. continuous (so ties occur with probability 0), or implement an explicit tie-breaking rule (e.g. add an independent continuous jitter, or use a strict filter  $F(C_i; \tau) = \{c \in C_i : s_\theta(P_i, R_i, c) > \tau\}$ ). Under either convention,  $S(C_i, W_i) = \max\{s_\theta(P_i, R_i, c) : W_i(c) = 0\}$  with  $\max \emptyset := -\infty$ .

## 2.2 Conditional conformal backbone

[Cherian et al.](#) build on conditional conformal prediction with function-class conditional guarantees [[Gibbs et al., 2024](#)]. The core idea is to replace unattainable pointwise conditional validity (impossible in general [[Barber et al., 2019](#)]) with linear calibration constraints indexed by a function class  $\mathcal{F}$ .

Let  $X_i = X(P_i, R_i)$  be features derived from the prompt and response (topic indicators, length, popularity proxies, etc.). Let  $S_i$  be a scalar conformity score induced by the filtering loss (e.g., (6) or, for a general monotone loss,  $S(C_i, W_i) = \inf\{\tau : L(F(C_i; \tau), W_i) \leq \lambda\}$  as in [Cherian et al. \[2024\]](#)). The conditional target in [Gibbs et al. \[2024\]](#) is stated for the *randomized conformal decision rule* produced by augmented quantile regression. Let  $\hat{\tau}_{\text{rand}}(\cdot)$  denote the randomized cutoff returned by their procedure and define the corresponding randomized filtered output

$$\hat{F}_{\text{rand}}(C_{n+1}) := F(C_{n+1}; \hat{\tau}_{\text{rand}}(X_{n+1})). \quad (7)$$

For the general-loss setting, define the success event for a new point as

$$E_{n+1} = \left\{ L(\hat{F}_{\text{rand}}(C_{n+1}), W_{n+1}) \leq \lambda \right\}. \quad (8)$$

The identity below is finite-sample and distribution-free under exchangeability, but it is *marginal over the random calibration sample* (and any internal randomization of the cutoff): conditioned on the realized calibration set, the left-hand side is generally a nonzero random quantity (which may concentrate near 0 as  $n$  grows). Their function-class conditional guarantee can then be written as the moment-calibration identity

$$\mathbb{E}\left[f(X_{n+1})(\mathbf{1}\{E_{n+1}\} - (1 - \alpha))\right] = 0, \quad \forall f \in \mathcal{F}. \quad (9)$$

If  $\mathcal{F}$  contains group indicators  $f(x) = \mathbf{1}\{x \in G\}$  with  $\mathbb{P}(X_{n+1} \in G) > 0$ , then (9) implies the finite-sample identity  $\mathbb{P}(E_{n+1} | X_{n+1} \in G) = 1 - \alpha$ , still marginal over the random calibration sample (and any internal randomization).

Technically, [Cherian et al. \[2024\]](#) obtain (9) via an *augmented quantile regression* construction: one solves a pinball-loss regression problem that includes an imputed test score as a decision variable, yielding a cutoff that satisfies the linear constraints exactly when a randomized version of the cutoff is used [[Gibbs et al., 2024](#)]. For linear  $\mathcal{F}$ , the computation reduces to a linear program, which later becomes important for differentiability.

**Contribution (i): conditional validity for general monotone losses.** A major conceptual step in [Cherian et al. \[2024\]](#) is to generalize from the single event “no false retained” to *any monotone loss* on filtered outputs. Let  $L(\widehat{F}(C_i), W_i)$  measure how bad the filtered output is relative to labels. They encode a high-probability constraint via the success event  $E_{n+1}$  in (8). When the constant function  $f \equiv 1$  belongs to  $\mathcal{F}$ , the moment identity (9) implies an *unconditional* (marginal) statement for the randomized rule:

$$\mathbb{P}\left(L(\widehat{F}_{\text{rand}}(C_{n+1}), W_{n+1}) \leq \lambda\right) = 1 - \alpha. \quad (10)$$

Without randomization (or with discrete scores and ties), one typically obtains a conservative inequality “ $\geq 1 - \alpha$ ” rather than equality.

To make conformal calibration possible, they assume two conditions. First, abstention is allowed:  $L(\emptyset, \cdot) = 0$ . Second,  $L$  is monotone in its first argument: if  $\widehat{F}_1 \subseteq \widehat{F}_2$ , then  $L(\widehat{F}_1, W) \leq L(\widehat{F}_2, W)$ . These assumptions ensure that the induced conformity score  $S(C_i, W_i) = \inf\{\tau \in \mathbb{R} : L(F(C_i; \tau), W_i) \leq \lambda\}$  is well-defined as an infimum (a generalized inverse); it is attained as a minimum under the same no-ties / tie-breaking convention used for Definition 3.

Their Theorem 3.1 establishes a *function-class conditional (moment) calibration identity* of the form (9) for the event  $\{L(\cdot) \leq \lambda\}$  (with a randomized cutoff for exactness), which implies group-conditional equalities when  $\mathcal{F}$  spans group indicators. A practical clarification is that the displayed guarantee is expressed as an expectation identity. It becomes an explicit probability statement only after choosing specific  $f$ 's (e.g., group indicators), and it remains marginal over the random calibration sample.

This general-loss view also clarifies the relationship to risk-control frameworks. Conformal risk control (CRC) targets control of an *expected* monotone loss [[Angelopoulos et al., 2025b](#)], while Cherian et al. target a *high-probability* constraint with conditional structure. Neither dominates the other. The choice depends on whether the application wants tail control (probability of exceeding a tolerance) or average control.

**Contribution (ii): level-adaptive conformal filtering.** Fixed  $\alpha$  can produce unusable outputs on hard prompts. [Cherian et al. \[2024\]](#) therefore allow an input-dependent reported level  $\alpha(X)$  and calibrate the reported level *in the same function-class conditional sense as before*. Let  $\widehat{\tau}_{l.a.,\text{rand}}(\cdot)$  denote the randomized cutoff returned by the level-adaptive procedure. Concretely, their Theorem 3.2 yields a moment-calibration identity for the Bernoulli event  $\{L(\cdot) \leq \lambda\}$ :

$$\mathbb{E}\left[f(X_{n+1})(\mathbf{1}\{L(F(C_{n+1}; \widehat{\tau}_{l.a.,\text{rand}}(X_{n+1})), W_{n+1}) \leq \lambda\} - (1 - \alpha(X_{n+1})))\right] = 0, \quad \forall f \in \mathcal{F}.$$

(with a randomized cutoff for exactness). This is *not* a pointwise conditional probability guarantee of the form  $\mathbb{P}(L \leq \lambda | X = x) \geq 1 - \alpha(x)$ ; it calibrates the issued levels over the slices encoded by  $\mathcal{F}$  (and, in particular, over bins of the reported  $\alpha(X)$  when such indicators are included in  $\mathcal{F}$ ).

In experiments,  $\alpha(\cdot)$  is learned on a separate fold to satisfy a utility criterion such as minimum claim retention. This learning step is outside the conformal guarantee and requires careful data splitting. Conceptually, level adaptivity changes what is being promised: the system no longer commits to a single global confidence level, but to a calibrated mapping from features to confidence. This idea resembles “many-slices” calibration: the paper highlights groupwise calibration over both feature groups and reported  $\alpha$  bins [[Jung et al., 2022](#)]. The function class  $\mathcal{F}$  mediates the trade-off between computational burden and how many slices are controlled.

**Contribution (iii): conditional boosting.** Even with conditional cutoffs, filtering can remove too many true claims when the score is weak. Cherian et al. propose *conditional boosting*: learn parameters  $\theta$  of a claim score  $p_\theta$  to maximize retained content on a hold-out set, while the conditional conformal constraint remains enforced.

They optimize an objective that counts how many claims survive the conformal threshold on held-out points. Because the objective contains indicators, they use a smooth sigmoid approximation as in prior marginal conformal training work [Stutz et al., 2021]. The main technical difficulty is that the conformal cutoff  $\hat{\tau}_i(\theta)$  depends on  $\theta$  through the calibration scores. For linear  $\mathcal{F}$ , the augmented quantile regression is a linear program, and the cutoff can be expressed via the *optimal basis* of the LP. Under a local stability condition (the optimal basis does not change under small perturbations), they derive a gradient expression by differentiating through this linear system (their Proposition 3.1).

This contribution is best viewed as a conditional analogue of learned/differentiable conformal methods [Bai et al., 2022, Einbinder et al., 2022]. The paper uses the gradient expression inside a training loop that repeatedly re-runs the conformal procedure on random splits, similar in spirit to conformal training algorithms [Stutz et al., 2021]. The key novelty is not just differentiating through calibration, but differentiating through calibration that enforces conditional constraints.

### 2.3 Experiments

Empirically, Cherian et al. [2024] show three consistent patterns. First, marginal conformal filtering can exhibit heterogeneous miscoverage across prompt/response slices (e.g., popularity strata in biographies), while the function-class conditional method stabilizes calibration over those slices. Second, level adaptivity improves utility by learning  $\alpha(X)$  to satisfy a retention target while preserving the same conditional calibration structure. Third, conditional boosting improves retention by learning a stronger claim score while re-enforcing the conditional conformal constraints during training. Full details and task-specific setups are in Appendix D.

**Limitations.** All guarantees are relative to the extraction+verifier pipeline that produces  $(C_i, W_i)$ , and require exchangeability of response-level units. These assumptions can fail under drift, feedback, or systematic verifier bias; Appendix A summarizes the scope conditions.

## 3 Future directions: response-level conformalization

Long-form answers contain many claims that are not independent. Claims share entities, premises, and local context. An early mistake can propagate as the model “explains” and justifies the mistake, a phenomenon known as hallucination snowballing [Zhang et al., 2023]. This creates within-response dependence in factual errors.

The selected paper already calibrates at the response level, i.e., each  $(P_i, R_i, C_i, W_i)$  is one exchangeable unit. Dependence mainly becomes a problem when practitioners (or downstream evaluations) revert to claim-level averaging or pseudo-replication. Claim-level metrics like FActScore average over claims [Min et al., 2023]. A claim-level error rate can look acceptable even when the probability that a *response* contains at least one error (or more than  $k$  errors) is high. Users consume responses, not independent claims, so reporting and optimization should target response-level risk.

It is tempting to “inflate” sample size by treating each extracted claim as an i.i.d. calibration or training unit (e.g., when fitting score parameters, learning  $\alpha(\cdot)$ , or reporting accuracy). This can break exchangeability because claims within a response are correlated. Even if one obtained a valid claim-level statement, it would not control response-level tail risk for long answers.

### 3.1 Formalization: response-level data unit and blocked conditional filtering

We adopt the response-level exchangeable unit (as in Cherian et al. [2024]) and make the “blocked” requirement explicit: all claims from a response must stay together whenever calibration, validation, or training splits are formed for any data-driven component.

**Definition 4 (Response-level exchangeable unit)** Let

$$Z_i = (P_i, R_i, C_i, W_i), \quad i = 1, \dots, n + 1,$$

be exchangeable response-level units, where  $P_i$  is the prompt (or prompt-side metadata),  $R_i$  is the generated response,  $C_i$  is the extracted set (or list) of claims, and  $W_i$  are claim-level labels.

We write  $W_i(c) \in \{0, 1\}$  for the label of claim  $c \in C_i$ , where  $W_i(c) = 1$  means “true” and  $W_i(c) = 0$  means “false” (according to the chosen verifier). Let  $\phi_i = \phi(P_i, R_i)$  be response-level features used for conditional calibration (e.g., claim count, response length, retrieval coverage proxy, or an entity-overlap proxy).

**Definition 5 (Response-level conditional filtering)** Given a claim score  $s_\theta(P_i, R_i, c)$  and a (possibly data-driven) threshold function  $\tau(\cdot)$ ,

$$F(C_i; \tau(\phi_i)) = \{c \in C_i : s_\theta(P_i, R_i, c) > \tau(\phi_i)\}. \quad (11)$$

We use strict thresholding ( $>$ ), unlike the standard ( $\geq$ ), to avoid ambiguity when multiple claims share the same score. This simplifies the definition of the conformity score  $S_k$  for discrete claim counts: setting  $\tau$  equal to the  $(k+1)$ -th false score allows us to exclude it.

Two assumptions make conformal calibration well-posed and align with the selected paper. First, the loss should be monotone in the filtered set (raising the threshold should not increase the loss). Second, abstention should be allowed: if  $F(C_i; \tau) = \emptyset$ , the loss is zero. In filtering, abstention corresponds to returning an empty answer or an explicit refusal.

**Choice of loss and conformity score.** To align the guarantee with user-facing risk, we propose a simple response-level tail loss that counts how many false claims remain after filtering.

**Definition 6 (Tail loss with fixed  $k$ )** Fix an integer  $k \geq 0$ . Define the response-level tail loss

$$L_k(F(C_i; \tau), W_i) = \mathbf{1}\left\{\left|\{c \in F(C_i; \tau) : W_i(c) = 0\}\right| > k\right\}. \quad (12)$$

This loss is monotone in  $\tau$  because increasing  $\tau$  can only remove claims. The case  $k = 0$  recovers the strict “no false retained” target [Mohri and Hashimoto, 2024]. Allowing  $k > 0$  can avoid vacuity when labels are noisy or when strict zero-error filtering collapses utility, but it must be justified for the application. The key benefit is interpretability: the guarantee is about a maximum tolerated number of factual errors in the displayed response. We then define a response-level conformity score that directly matches  $L_k$ .

**Definition 7 (Response-level conformity score for  $L_k$ )**

$$S_k(Z_i) = \inf \{\tau \in \mathbb{R} : L_k(F(C_i; \tau), W_i) = 0\}. \quad (13)$$

This score has an explicit characterization in terms of the scores of false claims, with tie-handling made explicit. Let the false claims in  $C_i$  have scores  $\{s_\theta(P_i, R_i, c) : W_i(c) = 0\}$  and sort them in descending order  $b_{i,1} \geq b_{i,2} \geq \dots$  (with the convention that the sequence is empty if there are no false claims). Then  $S_k(Z_i) = b_{i,k+1}$ , with the convention  $S_k(Z_i) = -\infty$  if fewer than  $k+1$  false claims exist.

This equality is exact under either of the following standard conventions: (i) strict filtering,  $F(C_i; \tau) = \{c \in C_i : s_\theta(P_i, R_i, c) > \tau\}$  (as in (11)); or (ii) non-strict filtering with an explicit tie-breaking rule (e.g., randomized jitter or lexicographic tie-breaking using an auxiliary  $U \sim \text{Unif}(0, 1)$ ). Under the strict rule, any threshold  $\tau \geq b_{i,k+1}$  guarantees that at most  $k$  false claims are retained, even when multiple false claims share the same score.

The advantage of blocked, response-level calibration is that it makes the exchangeability assumption more defensible and avoids pseudo-replication. It also yields a clean finite-sample statement.

**Proposition 1 (Response-level split conformal validity)** Assume  $(Z_i)_{i=1}^{n+1}$  are exchangeable response-level units as in Definition 4. Assume all data-dependent components of the factuality pipeline (claim extraction, verifier, and any learned parameters of  $s_\theta$ ) are fit on data disjoint from this calibration set, so that  $(S_k(Z_i))_{i=1}^{n+1}$  are exchangeable (or exchangeable conditional on the training fold). Compute response-level conformity scores  $(S_k(Z_i))_{i=1}^n$  on a calibration set of  $n$  responses, and let  $S_{(1)} \leq \dots \leq S_{(n)}$  be the sorted calibration scores. Define

$$m = \lceil (1 - \alpha)(n + 1) \rceil, \quad \hat{\tau} = S_{(m)},$$

with the standard convention  $S_{(n+1)} := +\infty$  (equivalently,  $\widehat{\tau}$  is the  $m/(n+1)$  quantile).

Then the filtered response produced by  $F(C_{n+1}; \widehat{\tau})$  satisfies the marginal, finite-sample guarantee

$$\mathbb{P}\left(L_k(F(C_{n+1}; \widehat{\tau}), W_{n+1}) = 0\right) \geq 1 - \alpha. \quad (14)$$

If the scores are continuous (or if ties are broken by adding an independent continuous jitter), the inequality is typically tight; otherwise the procedure is conservative. If one wants feature-dependent levels or slice-wise stability, one can apply a conditional conformal method to the scalar score  $S_k(Z_i)$  with covariates  $\phi_i$ . These methods yield  $\mathcal{F}$ -conditional calibration identities for the event  $\{L_k = 0\}$  (in the sense of moment constraints), rather than pointwise conditional guarantees.

This is standard split conformal applied to a response-level score. The contribution is the choice of exchangeable unit (one response), the blocked split, and a score/loss pair that targets a user-facing tail event. The guarantee remains conditional on the factuality pipeline (claim extraction and verifier) and on exchangeability at the response level. A response-level guarantee is not implied by claim-level control. Even strong claim-level statements can translate into weak user-facing safety when many claims are shown.

**Claim-level validity does not guarantee response-level safety.** Even if each retained claim has marginal error probability at most  $\alpha_c$ , the response-level error probability can be as large as  $\min\{1, m\alpha_c\}$  by a union bound when  $m$  claims are retained, and can vary widely depending on dependence. Appendix E states this formally.

### 3.2 Evaluation

We evaluate the proposed response-level tail-risk target by comparing (i) claim-level calibration with pseudo-replication and (ii) response-level blocked calibration based on  $S_k(Z_i)$  (Proposition 1). Primary metrics are the realized tail risk  $\mathbb{P}(L_k = 1)$ , retention (fraction of claims kept), and empty-response rate. To stress dependence, we stratify by claim count and by a simple dependence proxy such as named-entity overlap across claims. Ablations compare a constant cutoff  $\tau$  to  $\tau(\phi)$  using  $\phi = (k_i)$  and a small enriched feature set. Appendix F gives the full plan and robustness checks.

This direction complements Cherian et al. [2024]. Their framework already supports general monotone losses and conditional cutoffs. The proposed contribution is a *protocol* and a *reporting target*: treat responses as the exchangeable unit, enforce blocked calibration, and report response-level tail risks under dependence-aware stratifications. This helps avoid misleading interpretations that can arise from claim-level pseudo-replication and from averaging-based factuality metrics [Min et al., 2023, Zhang et al., 2023].

CRC and LTT can be layered on top if one wants to control other risk functionals. For example, CRC could control the expected number of retained false claims per response (a monotone loss), while conditional conformal tools could target stability across slices defined by  $\phi$  [Angelopoulos et al., 2025b,a, Gibbs et al., 2024]. The methodological choice should follow the intended user-facing statement.

**Limitations.** Response-level conformalization does not fix exchangeability violations, label noise, or unrepresentative benchmarks. Interactive deployment can break exchangeability through drift and feedback; online adaptive methods address some aspects but require continuous monitoring and reliable feedback signals [Gibbs and Candès, 2021]. Labeling pipelines (LLM judges, retrieval-based verifiers) can be wrong in systematic ways, making any guarantee only as good as the verifier [Min et al., 2023, Wei et al., 2024]. Finally, benchmarks like biographies and curated medical QA may not reflect all deployment contexts; conclusions should therefore be restricted to the tested distributions [Cherian et al., 2024, Manes et al., 2024].

## Conclusion

Cherian et al. [2024] advance conformal factuality filtering by providing conditional guarantees over meaningful slices, introducing level adaptivity to preserve utility on hard prompts, and proposing

conditional boosting to improve claim scores while respecting conditional conformal constraints. The broader lesson is that conformal guarantees are only as meaningful as (i) the conditioning structure one specifies, and (ii) the evaluation pipeline one adopts for factuality. As a future direction, response-level conformalization under dependent claims makes the exchangeable unit and the loss align with what users consume. Blocked calibration and response-level tail losses provide a clean finite-sample target and a practical evaluation protocol that can reveal dependence-driven failures that claim-level summaries may hide.

## References

- A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, abs/2110.01052, 2025a. URL <https://api.semanticscholar.org/CorpusID:238259034>.
- A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. *arXiv*, 2025b. URL <https://arxiv.org/abs/2208.02814>.
- A. Azaria and T. M. Mitchell. The internal state of an llm knows when its lying. *ArXiv*, abs/2304.13734, 2023. URL <https://api.semanticscholar.org/CorpusID:258352729>.
- Y. Bai, S. Mei, H. Wang, Y. Zhou, and C. Xiong. Efficient and differentiable conformal prediction with general function classes. In *International Conference on Learning Representations (ICLR)*, 2022. URL <https://arxiv.org/abs/2202.11091>.
- R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2019. URL <https://api.semanticscholar.org/CorpusID:88524668>.
- J. J. Cherian, I. Gibbs, and E. J. Candès. Large language model validity via enhanced conformal prediction methods. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.
- B.-S. Einbinder, Y. Romano, M. Sesia, and Y. Zhou. Training uncertainty-aware classifiers with conformalized deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL <https://arxiv.org/abs/2205.05878>.
- I. Gibbs and E. J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume abs/2106.00170, 2021. URL <https://api.semanticscholar.org/CorpusID:235266057>.
- I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024. URL <https://arxiv.org/abs/2305.12616>.
- M. Jeong, H. Hwang, C. Yoon, T. Lee, and J. Kang. Olaph: Improving factuality in biomedical long-form question answering. *ArXiv*, abs/2405.12701, 2024. URL <https://api.semanticscholar.org/CorpusID:269929793>.
- C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations (ICLR)*, volume abs/2209.15145, 2022. URL <https://api.semanticscholar.org/CorpusID:252668316>.
- S. Kadavath, T. Conerly, A. Askell, T. J. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Dodds, N. Dassarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. B. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022. URL <https://api.semanticscholar.org/CorpusID:250451161>.

- M. Kang, N. M. Gürel, N. Yu, D. Song, and B. Li. C-rag: certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024.
- B. Kumar, C.-C. Lu, G. Gupta, A. Palepu, D. R. Bellamy, R. Raskar, and A. L. Beam. Conformal prediction with large language models for multi-choice question answering. *ArXiv*, abs/2305.18404, 2023. URL <https://api.semanticscholar.org/CorpusID:258967849>.
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL <https://doi.org/10.1080/01621459.2017.1307116>.
- S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL <https://aclanthology.org/2022.acl-long.229/>.
- P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL <https://aclanthology.org/2023.emnlp-main.557/>.
- I. Manes, N. Ronn, D. Cohen, R. Ilan Ber, Z. Horowitz-Kugler, and G. Stanovsky. K-QA: A real-world medical Q&A benchmark. In D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.22. URL <https://aclanthology.org/2024.bionlp-1.22/>.
- S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL <https://aclanthology.org/2023.emnlp-main.741/>.
- C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024.
- H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.
- V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. Jaakkola, and R. Barzilay. Conformal language modeling. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 11654–11681, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/file/31421b112e5f7faf4fc577b74e45dab2-Paper-Conference.pdf](https://proceedings.iclr.cc/paper_files/paper/2024/file/31421b112e5f7faf4fc577b74e45dab2-Paper-Conference.pdf).
- G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008. URL <https://arxiv.org/abs/0706.3188>.
- D. Stutz, K. D. Dvijotham, A. T. Cemgil, and A. Doucet. Learning optimal conformal classifiers. *ArXiv*, abs/2110.09192, 2021. URL <https://api.semanticscholar.org/CorpusID:239016463>.
- V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 1st edition, 2005. Second edition published 2022.

J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, and Q. V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL <https://arxiv.org/abs/2403.18802>.

M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023. URL <https://arxiv.org/abs/2305.13534>.

## A Notation and scope assumptions

**Global notation.** A *prompt* is  $P$ , and an LLM produces a *response*  $R$ . A response is parsed into a list of  $k$  *claims*  $C = \{c_j\}_{j=1}^k$ . Each claim has a *truth label*  $W = \{w_j\}_{j=1}^k$ , with  $w_j = 1$  meaning “supported / factual” under the chosen verifier. A *claim score* is  $s_\theta(P, R, c)$  (or  $p(P, c)$ ), where larger values indicate higher confidence in the claim. Filtering keeps claims whose scores exceed a threshold.

**Scope.** Throughout, “validity” is always relative to a labeling mechanism. In factuality pipelines, labels can come from human annotation, from an LLM-as-judge, or from retrieval-augmented verification [Min et al., 2023, Wei et al., 2024]. A conformal guarantee is only as meaningful as the score and labels it wraps. All finite-sample conformal statements also require that the calibration and test units are exchangeable; this can fail under drift, feedback, or heavy prompt selection effects [Gibbs and Candès, 2021].

## B Other conformal interfaces for LLM outputs

Two broad conformal interfaces appear in LLM work. The first is *set generation*: output a set of strings that contains the true answer with high probability [Quach et al., 2024]. This aligns closely with classical conformal prediction, but it is awkward for many user interfaces because users want a single answer. The second is *filtering*: generate a response, decompose it into claims, and remove claims judged unreliable, producing a single filtered response [Mohri and Hashimoto, 2024]. Filtering is more user-friendly, but it moves the core difficulty into the factuality pipeline: claim extraction and claim scoring must be meaningful.

Certified risk control also appears at the pipeline level. For retrieval-augmented generation (RAG), C-RAG uses conformal risk control ideas to certify generation risks under assumptions on retrieval quality [Kang et al., 2024]. Discrete settings like multiple-choice QA provide a sanity-check environment where conformalization is straightforward and evaluation is cleaner [Kumar et al., 2023].

## C Factuality pipelines, benchmarks, and score signals (supplemental)

**Benchmarks (examples used in this paper).** Wikipedia biography factuality is a convenient controlled setting because claims can often be checked against a relatively stable corpus [Min et al., 2023]. Medical long-form QA is high-stakes and heterogeneous: OLAPH/MedLFQA aggregates multiple medical QA sources, and evaluation depends strongly on the chosen reference answers and citation practices [Jeong et al., 2024]; K-QA emphasizes real-world patient questions that differ from curated benchmarks [Manes et al., 2024]. These details are not needed for the conformal theory, but they matter for interpreting any empirical “validity” claim.

**Common score signals (non-exhaustive).** Claim scores can be built from behavioral uncertainty signals such as self-consistency across multiple samples [Manakul et al., 2023], from self-evaluation prompts, or from retrieval-based support signals [Min et al., 2023, Wei et al., 2024]. Work on self-knowledge suggests that models can sometimes predict when they know an answer in discrete QA [Kadavath et al., 2022], and representation-based probes can encode truthfulness-related signals [Azaria and Mitchell, 2023]. However, these signals can fail under *snowballing*: once a model commits to an incorrect premise, it may generate a consistent but wrong continuation [Zhang et al., 2023]. TruthfulQA highlights a distinct failure mode: confident imitation of misconceptions [Lin et al., 2022].

## D Experiments

The experiments in [Cherian et al. \[2024\]](#) support three empirical claims. First, conditional failures exist for marginal filtering. On Wikipedia biographies, the Mohri–Hashimoto split conformal method shows miscoverage varying across groups defined by page popularity (view counts), while the conditional method stabilizes miscoverage across those groups [[Cherian et al., 2024](#), [Mohri and Hashimoto, 2024](#)]. This matches the motivating concern: marginal validity can mask subgroup failures.

Second, level adaptivity improves utility when strict levels would be too conservative. In medical long-form QA (MedLFQA), the paper learns  $\alpha(X)$  to target a retention constraint (e.g., retain about 70% of claims), resulting in weaker levels on harder inputs while maintaining the conditional calibration structure [[Cherian et al., 2024](#), [Jeong et al., 2024](#), [Manes et al., 2024](#)]. This is an explicit and transparent trade: more content is kept, but the system reports a lower confidence level.

Third, conditional boosting improves retention relative to fixed scores. On synthetic heteroskedastic regression, conditional boosting avoids the conditional coverage failures that arise when optimizing only marginal constraints [[Cherian et al., 2024](#), [Stutz et al., 2021](#)]. On real LLM tasks, boosting combines multiple claim scores (frequency, self-evaluation, ordinal, etc.) to better correlate with labels and retain more text [[Cherian et al., 2024](#), [Mohri and Hashimoto, 2024](#)].

## E Claim-level control does not determine response-level risk

**Proposition 2 (Claim-level control does not determine response-level risk)** *Let  $E_j$  denote the event that the  $j$ -th retained claim in a response is false. A claim-level statement of the form  $\mathbb{P}(E_j) \leq \alpha_c$  does not determine the response-level error probability  $\mathbb{P}(\cup_{j=1}^m E_j)$  when  $m$  claims are retained. In general,*

$$\mathbb{P}\left(\bigcup_{j=1}^m E_j\right) \leq \min\{1, \sum_{j=1}^m \mathbb{P}(E_j)\} \leq \min\{1, m\alpha_c\}. \quad (15)$$

*If, for illustration,  $\mathbb{P}(E_j) = \alpha_c$  for all  $j$ , then  $\mathbb{P}(\cup_{j=1}^m E_j)$  can range from  $\alpha_c$  (perfect clustering of errors) up to  $\min\{1, m\alpha_c\}$  (nearly disjoint errors), and under independence it equals  $1 - (1 - \alpha_c)^m$ . Therefore, even a “small” per-claim error level can correspond to large (or highly variable) response-level risk when many claims are shown.*

The mismatch is exacerbated by positive dependence. In the presence of “snowballing” [[Zhang et al., 2023](#)], errors cluster within specific responses. While clustering might reduce the probability of *at least one* error compared to independence, it drastically increases the probability of *catastrophic failure* (many errors) once a failure occurs. Claim-level calibration is blind to this concentration risk; response-level calibration explicitly controls the tail of the error count distribution.

## F Experimental plan and ablations for response-level tail risk

We describe a concrete experimental plan. Fix a claim extraction pipeline (e.g., FActScore-style atomic facts [[Min et al., 2023](#)] or SAFE-style decomposition [[Wei et al., 2024](#)]) and fix a claim score family (e.g., frequency/self-evaluation/consistency signals as in [Mohri and Hashimoto \[2024\]](#), [Cherian et al. \[2024\]](#)). Then compare:

- **Claim-level calibration (pseudo-replication).** Treat each claim as a separate calibration point and calibrate a single threshold to control a claim-level loss.
- **Response-level blocked calibration.** Compute  $S_k(Z_i)$  per response and calibrate  $\hat{\tau}$  (or  $\hat{\tau}(\phi)$ ) using response-level units, as in Proposition 1.
- **Efficiency metrics.** For both methods, report the *Average Fraction of Claims Retained* and the *Empty Response Rate* (fraction of responses where filtering removes all claims).

Evaluate both methods on the response-level metric  $\mathbb{P}(L_k = 1)$  and on the distribution of the number of false retained claims.

**Dependence stress tests.** Stratify test responses by (a) claim count  $k_i$  and (b) a dependence proxy. One proxy is named-entity overlap across claims (high overlap suggests shared premises). Another proxy is a “snowballing score” based on repeated entity mentions following an initial unsupported claim, operationalized via verifier disagreement patterns [Zhang et al., 2023]. Report response-level tail risks across these strata.

**Conditional calibration and ablations.** Implement response-level conditional calibration using features  $\phi_i$  in  $\tau(\phi_i)$ , using the conditional conformal framework on  $S_k$  [Gibbs et al., 2024]. At minimum, include:

- **Ablation 1 (no features):**  $\tau(\phi) \equiv \tau$  (constant threshold).
- **Ablation 2 (claim count only):**  $\phi = (k_i)$ .
- **Full model:**  $\phi = (k_i, \text{length}, \text{retrieval coverage proxy}, \text{entity overlap})$ .

**Additional robustness checks.** Vary the generation temperature to change within-response diversity and error correlations. Swap the verifier/labeling mechanism (e.g., SAFE-style search vs. a self-consistency judge) to test sensitivity to label noise [Wei et al., 2024, Manakul et al., 2023]. Include at least one dataset outside biographies and MedLFQA, such as TruthfulQA prompts, to test imitative falsehood settings [Lin et al., 2022].

## G Response-level score and mismatch examples

### Response-level conformity score details

For the tail loss  $L_k$  in (12), define the multiset of false-claim scores in response  $i$ :

$$\mathcal{B}_i = \{s_\theta(P_i, R_i, c) : c \in C_i, W_i(c) = 0\}. \quad (16)$$

Let  $b_{i,1} \geq b_{i,2} \geq \dots$  be the sorted elements of  $\mathcal{B}_i$  in descending order. Consider the strict filtering rule (11) where a claim is retained if  $s > \tau$ . To ensure that *at most*  $k$  false claims are retained, we must filter out the  $(k+1)$ -th largest false score and all smaller scores. Setting  $\tau = b_{i,k+1}$  achieves this, as any claim with score  $s > b_{i,k+1}$  corresponds to indices  $1, \dots, k$ . Thus, the conformity score is exactly  $S_k(Z_i) = b_{i,k+1}$  (with  $S_k = -\infty$  if  $|\mathcal{B}_i| \leq k$ ).

### A short mismatch counterexample (independent claims)

Consider a claim-level procedure that ensures each retained claim is false with probability  $\alpha_c$ , and suppose a response contains  $m$  retained claims with independent errors. Then the probability the response is fully correct is  $(1 - \alpha_c)^m$ , so the probability of at least one error is  $1 - (1 - \alpha_c)^m$ . For  $\alpha_c = 0.01$  and  $m = 200$ , this is  $1 - (0.99)^{200} \approx 0.866$ . Thus, even “1% per-claim” can correspond to an 86.6% chance of at least one error at the response level. This illustrates why response-level calibration is the appropriate default when the user-facing unit is a response.