Multi-group Uncertainty Quantification for Long-form Text Generation

Terrance Liu[1] **Zhiwei Steven Wu**[1]

[1]Carnegie Mellon University

Abstract

While past works have shown how uncertainty quantification can be applied to large language model (LLM) outputs, the question of whether resulting uncertainty guarantees still hold within sub-groupings of data remains open. In our work, given some long-form text generated by an LLM, we study uncertainty at both the level of individual claims contained within the output (via *calibration*) and across the entire output itself (via *conformal prediction*). Using biography generation as a testbed for this study, we derive a set of (demographic) attributes (e.g., whether some text describes a man or woman) for each generation to form such "subgroups" of data. We find that although canonical methods for both types of uncertainty quantification perform well when measuring across the entire dataset, such guarantees break down when examining particular subgroups. Having established this issue, we invoke group-conditional methods for uncertainty quantification—*multicalibration* and *multivalid* conformal prediction—and find that across a variety of approaches, additional subgroup information consistently improves calibration and conformal prediction within subgroups (while crucially retaining guarantees across the entire dataset). As the problems of calibration, conformal prediction, and their multi-group counterparts have not been extensively explored in the context of long-form text generation, we consider these results to form a benchmark for this setting.

1 INTRODUCTION

In recent years, researchers have developed stronger large language models that perform well on a variety of tasks across different domains [Touvron et al., 2023, Bubeck et al., 2023, Anil et al., 2023]. However, as use of LLMs continues to grow, so do concerns over their tendency to hallucinate facts [Huang et al., 2023]. As a result, there is a growing need for methods that can reduce hallucinations [Manakul et al., 2023, Zhang et al., 2023], perform abstention [Yang et al., 2023], or provide correctness guarantees [Kumar et al., 2023, Mohri and Hashimoto, 2024, Quach et al., 2023]. Our work focuses on the latter—broadly speaking, uncertainty quantification of long-form large language model generations.

Concretely, given a set of claims produced by an LLM in response to some prompt, our goal is to provide a confidence score or uncertainty guarantee about the factual correctness of the output. We explore this problem in two settings: given a set of claims contained within some long-form prompt response, we **(1)** ensure factuality at the individual claim level and **(2)** provide uncertainty guarantees across the whole set of claims. We approach problem **(1)** via *calibration*, in which one wishes to output a calibrated score for each claim, while for problem **(2)**, we apply *conformal prediction* [Shafer and Vovk, 2008], selecting a subset of claims that—with high probability—are *all* correct.

In contrast to existing works on uncertainty guarantees of long-form generations [Quach et al., 2023, Mohri and Hashimoto, 2024], we make the observation that while these guarantees may be valid under the full data distribution, they may not still be valid within individual subgroups of the distribution. For example, generations describing local politicians may be more prone to error than generations concerning national leaders. We choose biography generation as a testbed for multi-group uncertainty quantification, arguing that this problem is well-motivated, given that bias within biography generation has long been studied [De-Arteaga et al., 2019]. Having derived a set of subgroups using demographic information (e.g., whether an LLM output describes a man or woman), we find that when evaluated with respect to such groupings, canonical methods for calibration and

conformal prediction indeed exhibit significant biases.[1]

Having established such issues for standard uncertainty quantification approaches, we shift our attention to understanding to what extent such biases can be corrected. To address this unmet need, we introduce methods quantifying uncertainty in long-form text generation that are valid not only across a full distribution of prompts (i.e., marginally) but also across identifiable subgroups of prompts (i.e., conditionally). Invoking **(1)** *multi*calibration [Hébert-Johnson et al., 2018] and **(2)** *multivalid* conformal prediction [Jung et al., 2022], we categorize methods into two styles: iterative "patching" and linear regressor-based algorithms.

Our results demonstrate that for both problems **(1)** and **(2)**, multicalibration and multivalid conformal prediction techniques improve measures of uncertainty relative to standard (marginal) calibration and conformal prediction methods. This advantage holds ***regardless*** of whether evaluation is conducted within groups or across the entire dataset. As the problems of calibration, conformal prediction, and their multi-group counterparts have not been extensively explored in the context of long-form text generation, we consider these results to form a benchmark for this setting.

1.1 RELATED WORK

Factuality in long-form LLM outputs. Evaluating factuality for long-form generation [Min et al., 2023a, Song et al., 2024, Wei et al., 2024, Bayat et al., 2024] is challenging: not only do generated outputs consist of many parts that must be scored individually, but also scoring each part requires prohibitively costly manual annotation. To make evaluation more tractable, Min et al. [2023a] introduce FACTSCORE, which converts any generation into a set of atomic facts (claims) that are then labeled as true or false. Using this evaluation metric, Min et al. [2023a] test LLMs' abilities to generate biographies and find that their generations are pervaded with errors.

Attaching confidence scores to LLM outputs. While a natural method for producing an uncertainty estimate is to use a model's output probabilities directly as a confidence score [Achiam et al., 2023], it has been shown that model probabilities are not well calibrated [Guo et al., 2017]. As a result, many works have recently proposed alternative methods for generating uncertainty scores that can then be used to refine or correct LLM outputs [Wang et al., 2022, Xiong et al., 2023, Geng et al., 2024, Fadeeva et al., 2023, Vashurin et al., 2024]. We highlight that such work is complementary to our line of work—rather than proposing an entirely new uncertainty score function, we focus on how one can better leverage existing scores to produce uncertainty guarantees.

Uncertainty quantification for LLMs. We note that much of the prior work on multicalibration and multivalid conformal prediction are rooted in theory. Like Detommaso et al. [2024], our work tries to bridge the gap between theoretical insights and practical problems today (i.e., LLM generations). However, while Detommaso et al. [2024] calibrate for correctness in question-answering, we are the first to apply multicalibration to claims decomposed from long-form text generation. Moreover, unlike Detommaso et al. [2024], we consider uncertainty quantification in the form of conformal prediction.

In addition, our work closely relates to Mohri and Hashimoto [2024], which aims to provide high probability guarantees of factuality in long-form generation. In particular, Mohri and Hashimoto [2024] frame this problem as a nested conformal prediction problem, producing subsets of claims that achieve some marginally valid coverage (i.e., produce some generation that on average, contains a correct output with any user-specified probability). Our work, however, extends this problem to multivalid conformal prediction: we produce generations that are not only correct on average but are also conditionally correct across subgroups.

Finally, concurrent work by Cherian et al. [2024] also builds off this framework, but unlike our work, they introduce a new objective in which the goal is to instead guarantee that at least some given proportion of claims are retained. By applying a method proposed in Gibbs et al. [2023], Cherian et al. [2024]'s algorithm can (optionally) condition on group membership. However, their experiments include only 5 (non-overlapping) groups that are derived from the same feature, while our work focuses on the more challenging setting in which examples can simultaneously belong to many groups.

2 PRELIMINARIES

2.1 CALIBRATION

We begin by defining calibration in context of factuality in open-ended text generation. Suppose we are given some $(X, Y) \sim \mathcal{D}$ where $X \in \mathcal{X}$ denotes some claim outputted by an LLM, while Y is an indicator in which $Y = 1$ when the claim is correct (and $Y = 0$ otherwise). Suppose there exists some uncertainty score function $f : \mathcal{X} \to [0, 1]$ that measures confidence for the correctness of some input X (with higher values denoting higher levels of confidence). Then a goal one may have when designing such a score function f is to have that

$$P_{\mathcal{D}}(Y = 1 \mid f(X) = p) = p, \forall x \in \mathcal{X} \tag{1}$$

In other words, the probability that some LLM output is correct is given exactly by f.

[1] Uncertainty can be epistemic and aleatoric, and sources of group bias can be categorized into either (or both) types. Our work, which focuses on atomic factuality in long-form generation, falls under epistemic uncertainty.

Calibration, then, defines a simpler, more tractable condition, in which instead of ensuring guarantees across all possible values of f, it ensures a guarantee over coarser, level sets $S_p(f)$:

Definition 1. *(Calibration) A function f is calibrated w.r.t D if*

$$\Delta_p(f) = 0, \forall p \in [0, 1]$$

where $\Delta_p(f)$ is the bias of f for the p-th level set $S_p(f) = \{f(x) = p\}$:

$$\Delta_p(f) = \mathbb{E}_{\mathcal{D}}[Y - f(X) \mid S_p(f)]$$

Defining level sets is akin to dividing the output space of f (i.e., $[0, 1]$) into buckets. For example, one could round $f(X)$ to the nearest value in some predefined set of probabilities (e.g. $\{0, 0.5, 1.0\}$). One can view this definition of calibration as a desirable guarantee since it serves as a minimal condition for Equation 1—any f that satisfies (1) must (at the very least) also be calibrated. We note that to evaluate calibration, we can consider the average squared calibration error (ASCE) of f.

$$\text{ASCE}(f) = \mathbb{E}_P[\Delta_P^2(f)] \tag{2}$$

The ASCE averages the squared bias across all level sets and is zero when f is calibrated.

Multicalibration. While calibration provides an already important and useful guarantee, it can often be insufficient in many real-world scenarios. For example, in the context of generating information about people, one maybe desire that f is calibrated not only across all people, but also within subpopulations defined by demographic attributes like *sex or gender*. Otherwise, it is possible that certain subgroups can still suffer from very high miscalibration, even when the score function is perfectly calibrated across $\mathcal{D}$. Ideally, one would hope to have guarantees while conditioning on as many subgroups in $\mathcal{X}$ as possible, both from the perspective of machine learning fairness as well as enhancing the likelihood of correctness in general.

Multicalibration [Hébert-Johnson et al., 2018] was developed to provide accurate guarantees across overlapping subgroups (i.e., a sample can belong to many groups). Let $g : \mathcal{X} \to \{0, 1\}$ be a group function that evaluates to 1 if X belongs to some group. We study, then, the setting in which there exists of set of groups $\mathcal{G}$ that corresponds to our data domain $\mathcal{D}$. While the set of groups can be disjoint, the problem of multicalibration then becomes trivial in this case because one can simply split a dataset into disjoint sets that can then each be calibrated individually. Consequently, prior work typically considers the more interesting case where many intersecting groups comprise $\mathcal{G}$.

Given a group function g, we define group average squared calibration error (gASCE) as:

$$\text{gASCE}(f, g) = \mathbb{E}_P[\Delta_{p,g}^2(f) \mid g(X) = 1] \tag{3}$$

where

$$\Delta_{p,g}(f) = \mathbb{E}_{\mathcal{D}}[Y - f(X) \mid S_{p,g}(f)]$$

for $S_{p,g} = \{f(X) = p, g(x) = 1\}$. In other words, gASCE conditions on both level sets and group membership. Finally, we have:

Definition 2. *(Multicalibration) A function f is α-multicalibrated w.r.t D and a set of groups $\mathcal{G}$ if and only if*

$$gASCE(f, g) < \frac{\alpha}{P_{\mathcal{D}}(g(X) = 1)}, \forall g \in \mathcal{G}$$

2.2 CONFORMAL PREDICTION

In conformal prediction, the general goal is to produce some confidence set $\mathcal{T}(X)$ for some example X such that this set marginally *covers* the true label Y with some target probability $1 - \alpha$.

$$P_{\mathcal{D}}(Y \in \mathcal{T}(X)) = 1 - \alpha \tag{4}$$

The second part of our work follows the problem statement outlined in Mohri and Hashimoto [2024]. Unlike in calibration, where each claim contained in some long-form generation is treated individually, Mohri and Hashimoto [2024] instead define their problem in terms of pairs (X, Y), where X is some input prompt and $L(X) = Y \in \mathcal{Y}$ is the long-form generation outputted by a LLM L. Because Y may or may not be supported by some reference ground truth Y^*,[2] Mohri and Hashimoto [2024] define factuality in terms of entailment operations $Y^* \implies Y$. Furthermore, they rewrite this relation as $Y^* \in E(Y) = \{Y' \in \mathcal{Y} : Y' \implies Y\}$. This equivalent set notation, in other words, means that some reference ground truth Y^* (e.g., a Wikipedia article in Min et al. [2023a]) is contained in the set of possible texts Y' that support all claims made in the LLM output Y.

Given this notation, the goal is to find some uncertainty set $\mathcal{T}(L(X))$ s.t. $P_{\mathcal{D}}(Y \in \mathcal{T}(L(X))) = 1 - \alpha$. In the context of long-form text generation, this goal translates to taking as input the original LLM output $L(X)$ and producing a subset of claims $\mathcal{T}(L(X))$ such that with high probability, $1 - \alpha$, all remaining claims are factually correct.

We note that to empirically measure such guarantees, one can use the *coverage error* of $\mathcal{T}$ w.r.t the target error rate α.

$$|P_{\mathcal{D}}(Y \notin \mathcal{T}(X)) - \alpha| \tag{5}$$

Multivalid Conformal Prediction. Similar to calibration, one may also desire group conditional coverage guarantees for intersecting groups. Known as *multivalid conformal prediction* [Jung et al., 2022], these guarantees are stronger

[2] in the case of FActScore [Min et al., 2023a], "is Y supported by Wikipedia?"

than marginal conformal guarantees, holding also when conditioned on group membership. Using group functions g, as defined in Section 2.1, full multivalid coverage can be written as the following: Given some set of groups $\mathcal{G}$, we have that

$$P_{\mathcal{D}}(Y \in \mathcal{T}(X) \mid g(X) = 1) = 1 - \alpha \tag{6}$$

for all group functions $g \in \mathcal{G}$. Thus, target coverage guarantees $1 - \alpha$ must hold both marginally and within all subgroups.

3 METHODS

Next, we introduce the methods (and their group-conditional variants) for applying calibration and conformal prediction to language model factuality. We organize these methods into two categories: **(1)** iterative "patching"-based algorithms and **(2)** linear regressor algorithms. As mentioned previously, prior exploration of long-form text generation has been limited. While Mohri and Hashimoto [2024] evaluate one variant—split conformal (SC)—on a small set of entities, we are not aware of prior work that has considered other uncertainty quantification methods in this setting.

3.1 ITERATIVE "PATCHING" ALGORITHMS

The first category of algorithms can be characterized as patching algorithms. Given a base method for calibration or conformal prediction, one iterates through groups $g \in \mathcal{G}$ in which the method does poorly on. At each iteration, the algorithm corrects the bias (i.e., patches up the function) on just that subset of examples (i.e., $g(x) = 1$). Once some stopping condition is met,[3] the final, "patched up" function satisfies multi-group guarantees.

Calibration. For calibration, we consider *Histogram Binning* (HB) [Zadrozny and Elkan, 2001], presented in Algorithm 1. This method, takes some base scoring function f and discretizes the output space to a set of p-th level sets $S_p(f)$, as defined in Section 2.1. Given some target grid of values $p \in [\frac{1}{m}]$, we round f to the closest value in the grid

$$f'(x) = \underset{p \in [\frac{1}{m}]}{\operatorname{argmin}} |f(x) - p|.$$

[3] In the standard formulation of iterative patching, the stopping criteria is set as a function over the number of bins so that one can prove guarantees about algorithm (see Roth [2022]). In practice, we found this stopping criteria to be too conservative, and so we instead run iterative patching on the calibration and test sets concurrently and use the calibration set to determine the stopping iteration (i.e., we enforce early stopping once we can no longer make improvements on the calibration set).

Algorithm 1 Histogram Binning (HB)

Input: scoring function f'
for $p \in [\frac{1}{m}]$ **do**
  Set

$$\hat{f}(x) = \begin{cases} f'(x) + \Delta_p(f') & \text{if } x \in S_p(f') \\ f'(x) & \text{otherwise} \end{cases}$$

end for
Output: $\hat{f}$

Algorithm 1 then applies a constant correction[4] for each level set $S_p(f)$ in the grid, based on the calibration error of the model f'.

In Algorithm 2, we present the multi-group version of histogram binning, known as *Iterative Grouped Histogram Binning* (IGHB) [Hébert-Johnson et al., 2018]. In this algorithm, we instead apply a constant correction conditioned on $S_{p,g}$ (i.e., both the level set and group membership). At each step t, IGHB identifies $S_{p,g}$ for which the calibration error (weighted by the group size) is highest and then corrects it for this level set *and* group. The algorithm then continues until some stopping condition is met, iteratively patching f' for various groups $g \in \mathcal{G}$.

Algorithm 2 Iterative Grouped Histogram Binning (IGHB)

Input: scoring function f', max iterations T
Let $H_t(p, g) = P_D(S_{p,g}(f_t))\Delta^2_{p,g}(f_t)$
Initialize $f_0 = f'$
for $t \in \{0, 1, \ldots, T-1\}$ **do**
  Set

$$(p_t, g_t) = \underset{p \in [\frac{1}{m}], g \in \mathcal{G}}{\operatorname{argmax}} H_t(p, g)$$

  Let $\Delta_t = \Delta_{p_t, g_t}$ and $S_t = S_{p_t, g_t}$
  Set

$$h_{t+1}(x) = \begin{cases} f_t(x) + \Delta_t(f_t) & \text{if } x \in S_t(f_t) \\ f_t(x) & \text{otherwise} \end{cases}$$

  Set $f_{t+1} = h_{t+1}$
  if $H_t(p_t, g_t) \geq H_{t-1}(p_{t-1}, g_{t-1})$ **then**
    Set t = t-1
    break
  end if
end for
Output: H_t

[4] In Algorithms 1 and 2, we assume true data distribution is given, and therefore we can calculate $\Delta_{p,g}$. In practice (and our experiments) $\Delta_{p,g}$ is estimated using a calibration set.

Conformal prediction. We first present the *Split Conformal* (SC) method [Shafer and Vovk, 2008, Gupta et al., 2022]. In particular, we consider the standard approach where one constructs a set of nested sets and each output set contains some subset $\mathcal{F}(\mathcal{X})_t$ of claims generated by the LLM.

Following Mohri and Hashimoto [2024], we define these nested sets $\mathcal{T}$ as thresholds sets where each set $\mathcal{F}(L(X))$ contains the set of all individual claims $\{x \in L(X) \mid f(x) > t\}$ for some scoring function f. More formally, we have that $\mathcal{F}(L(X))_{t\in\mathcal{T}}$ satisfies the nested sequence property if for $t, t' \in \mathcal{T}, t \leq t'$, we have that $\mathcal{F}_t(L(X)) \subseteq \mathcal{F}_{t'}(L(X))$.

To construct these threshold sets, we have that

$$r(X, Y) = \inf\{t \in \mathcal{T}, Y \in \mathcal{F}_t(L(X))\}$$

where r defines the minimum safe threshold such that $Y \in \mathcal{F}_t(L(X))$ for all $t > r(X, Y)$. Practically speaking, given some set of uncertainty scores $f(x)$ for each claim $x \in L(X)$, $r(X, Y)$ defines the minimum value such that any set of claims $\mathcal{F}_t(L(X)) = \{x \in L(X) \mid f(x) \geq t\}$ will be entirely true if and only if $t \geq r(X, Y)$.

Given some calibration set $\hat{D}$ of size n and some target error rate α (or target coverage $1 - \alpha$), split conformal simply outputs the set $\mathcal{F}_{q_\alpha}(L(X))$ for any X, where q_α is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th-quantile of scores $\{r(X_i, Y_i)\}_{i=1}^n$ for $X_i, Y_i \in \hat{D}$.

In Algorithm 3, we present the *multivalid split conformal* (MVSC) prediction technique that closely resembles methods originally proposed in Jung et al. [2022]. Similar to IGHB, we start with some base threshold (i.e., the threshold q_α obtained from using split conformal). Then at each iteration t, we find the group g_t that has the worst squared coverage error $\Delta_{t,g}$, weighted by the size of the group $P(g_t(X) = 1)$. Then, we simply "patch" the thresholds for examples $\{(X, Y) \mid g_t(X) = 1\}$, again using the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th-quantile of scores for (X, Y) belong to group g_t. Like in IGHB, we continue patching the set of thresholds until some stopping criterion is met.

3.2 LINEAR REGRESSOR ALGORITHMS

Next, we consider algorithms that instead solve an optimization problem for the purpose of calibration and conformal prediction. In these cases, one can naturally make them multi-group/valid by including group-membership (i.e., $g(X) = 1$ for all $g \in \mathcal{G}$) in the optimization problem itself. Formally, we describe these linear regression based methods in Algorithms 4 and 5. Presented in this way, the methods for calibration vs. conformal prediction is reduced to a choice of loss function L. Again, we assume one has access to some calibration set for which one solves the optimization problem on.

Algorithm 3 Multivalid Split Conformal (MVSC)

Input: calibration set $\hat{D}$, LLM L, fact-level scoring function f, target error rate α, split conformal threshold q_α, max iterations T
Let $\mathcal{F}_{h_t}(L(X)) = \{x \in L(X) \mid f(x) \geq h_t(X)\}$
Let $\Delta_{t,g} = P_{\mathcal{D}}(Y \in \mathcal{F}_{h_t}(L(X)) \mid g(X) = 1)$
Let $H_t(g) = P_{\mathcal{D}}(g(X) = 1)[(1 - \alpha) - \Delta_{t,g}]^2$
Initialize $h_0(X) = q_\alpha$
for $t \in \{0, 1, \ldots, T - 1\}$ **do**
  Set

$$g_t = \operatorname*{argmax}_{g\in\mathcal{G}} H_t(g)$$

  Let $\hat{D}_t = \{(X, Y) \in \hat{D} \mid g_t(X) = 1\}$
  Set q_t to be the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th-quantile of scores $\{r(X_i, Y_i)\}$ for $X_i, Y_i \in \hat{D}_t$
  Set

$$h_{t+1}(X) = \begin{cases} q_t & \text{if } g_t(X) = 1 \\ f_t(X) & \text{otherwise} \end{cases}$$

  if $H_t(g_t) \geq H_{t-1}(g_{t-1})$ **then**
    Set t = t-1
    break
  end if
end for
Output: h_t

Calibration. For calibration, one can choose L to be binary cross-entropy loss. In doing so, Algorithm 4 then describes *Platt Scaling* (PS) [Platt, 1999], which can be described as fitting a logistic regression model to some set of model outputs to obtain calibrated probability scores.[5] Algorithm 5 describes the multi-calibrated version of Platt Scaling. While not explicitly derived in their work, this multicalibration formulation can be traced back to Gopalan et al. [2022], who establish a hierarchy of notions for multicalibration and analyze multicalibration on functions trained with linear loss. Going forward, we refer to this method as *Group Conditional Unbiased Logistic Regression* (GCULR).

Conformal prediction. For conformal prediction, we instead choose L to be pinball loss. We refer to the non-group version of this method (Algorithm 4) as *Conformalized Quantile Regression* (CQR) [Romano et al., 2019], in which

[5] A related calibration method to Platt scaling (PS) is temperature scaling (TS) [Guo et al., 2017], which was originally introduced for calibrating neural networks for multiclass classification and has been incorporated in work on calibrating NLP models [Sicilia et al., 2024]. We note, however, that in the binary classification setting (e.g., our setting where we identify if an output is correct or not), TS is mathematically equivalent to PS when there is no bias term and the weight takes on the form $\frac{1}{\tau}$, where τ is the temperature learned in TS.

Algorithm 4 Linear Regressor

1: **Input:** data distribution $\mathcal{D}$, scoring function f, loss function L
2: Set

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \, \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[L\left(f(X;\lambda), Y \right) \right]$$

s.t. $f(X;\lambda) = \lambda_0 + \lambda_1 f(X)$
3: **Output:** $f(X;\hat{\lambda})$

Algorithm 5 Group-conditional Linear Regressor

1: **Input:** data distribution $\mathcal{D}$, scoring function f, loss function L, set of groups $\mathcal{G}$
2: Set

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \, \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[L\left(f(X;\lambda), Y \right) \right]$$

s.t. $f(X;\lambda) = \lambda_0 + \lambda_1 f(X) + \sum_{\lambda_g \in \mathcal{G}} \lambda_g g(x)$
3: **Output:** $f(X;\hat{\lambda})$

given some target coverage $1-\alpha$, we fit a linear quantile regression model that minimizes pinball loss.

In our conformal prediction setting, as described in Section 2.2, X is an entire biography, or set of independent claims. Thus, to adapt quantile regression to long-form generation, we propose setting $f(X)$ to be a vector of uncertainty scores for each claim $x \in X$. Like in split conformal, the target is then the minimum threshold $r(X,Y)$ for which all claims above it are correct. In the multivalid case, we then add group features $g(X)$ to the optimization problem. A version of Algorithm 5 was first presented by Jung et al. [2022], and going forward, we will refer to this method as *Group Conditional Conformalized Quantile Regression* (GCCQR).

We note that in our experiments, each biography generated by the LLM may have a different number of claims, a setting in which prior work on conformal quantile regression does not account for. Consequently, we propose using interpolation to (un)squeeze the set of scores to a vector $f(X)$ of fixed size ($K = 25$ in our experiments). While Mohri and Hashimoto [2024] only show that split conformal can be applied to this type of setting, our experiments demonstrate that quantile regression methods achieve similar performance for marginal (CQR) and multigroup (GCCQR) methods (Section 5).

4 EMPIRICAL EVALUATION

We focus our empirical evaluation on the problem of biography generation, which we contend serves as a very suitable testbed for evaluating factuality and has been used as a benchmark in a variety of works in recent years. Outputting biographies offers one the ability to evaluate not only a set of objective and specific claims but also on a wide range of topics, which in turn allows us to explore a rich set of group functions for each person. Moreover, bias within biography generation has long been a studied issue, further motivating the problem of ensuring group-conditional guarantees. Like Min et al. [2023a], we use a language model to automate the process of decomposing biographies into claims and evaluating for factuality (Appendix B.1).

Dataset We evaluate on a large set of biographies by extracting 8,541 entities from the Natural Questions dataset [Kwiatkowski et al., 2019], which consists of real queries issued to the Google search engine. We denote this dataset as BIO-NQ. Our motivation for choosing Natural Questions is that these extracted human entities should serve as a representative sample of public figures that users may prompt an LLM about. For each question, we select all entities in either the question's short answer or accompanying Wikipedia article. We then attempt to match them to their corresponding Wikidata entry. If a match exists and its Wikidata page's property, *if instance of*, is equal to the value, *human*, we add the entity to our dataset BIO-NQ.

Collecting group features To obtain groups for each person found in our dataset, we extract properties by scraping Wikidata for each entity and identifying ones that are commonly shared among entities in BIO-NQ. The exact group attributes we use in our experiments are described in Appendix B. To form groups $\mathcal{G}$ from these attributes, we take all 1 and 2-way combinations of attributes and the values they take on, giving us $|\mathcal{G}| = 77$ subgroups.

Generating confidence scores The algorithms described in Section 3 require a base scoring function. For experiments, we use the following:

1. **Self-consistency** [Wang et al., 2022]: Our first score is a frequency-based scoring function inspired by *self-consistency*. To score each claim found in a generated biography, we prompt the LLM to output a biography M additional times. We use the proportion of times the claim is contained in the additional reference generations as the uncertainty score. We automate the calculation of this score using BM25 and AlignScore [Zha et al., 2023] (See Appendix B.2).
2. **P(True)** [Kadavath et al., 2022]: For each biographical claim, we prompt the LLM to assess whether it is true or false. We then output the ratio of next token probabilities of the tokens for "true" and "false": $\frac{P(True)}{P(True)+P(False)}$.
3. **Verbalized confidence** [Tian et al., 2023]: To output *verbalized confidence* as an uncertainty estimate, one prompts the LLM to directly output their confidence level in its response. We originally tried having the

Model	Base Score	Metric	Uncalibrated	HB	IGHB	PS	GCULR
Llama	self-consistency	marginal	0.32291	0.00875	**0.00038**	0.00022	**0.00015***
		group max	0.42343	0.07711	**0.01481**	0.05791	**0.00628***
		group mean	0.33352	0.01597	**0.00289**	0.00636	**0.00111***
	P(True)	marginal	0.11768	0.00451	**0.00021***	0.00036	**0.00022**
		group max	0.20701	0.06988	**0.01798**	0.06025	**0.00697***
		group mean	0.12682	0.01176	**0.00328**	0.00654	**0.00145***
	verb. conf.	marginal	0.01642	**0.00014**	0.00055	**0.00013***	0.00023
		group max	0.06645	0.06634	**0.01315**	0.06709	**0.00730***
		group mean	0.02447	0.00738	**0.00357**	0.00750	**0.00154***
Mistral	self-consistency	marginal	0.30706	0.01163	**0.00026**	0.00029	**0.00013***
		group max	0.43659	0.08372	**0.01660**	0.05729	**0.00487***
		group mean	0.32067	0.01988	**0.00269**	0.00726	**0.00097***
	P(True)	marginal	0.06291	0.00074	**0.00031**	0.00047	**0.00015***
		group max	0.12293	0.06942	**0.01417**	0.07054	**0.00587***
		group mean	0.07173	0.00896	**0.00309**	0.00886	**0.00132***
	verb. conf.	marginal	0.22229	0.00047	**0.00036**	0.00034	**0.00015***
		group max	0.33763	0.06922	**0.01453**	0.07080	**0.00653***
		group mean	0.23230	0.00869	**0.00315**	0.00878	**0.00128***

Table 1: We generate biographies using Llama 2 7B Chat and Mistral 7B Instruct for entities from BIO-NQ and compare each calibration method (HB, PS) against its multicalibration counterpart (IGHB, GCULR) on **ASCE**, **max gASCE**, and **average gASCE** (↓ better). We test each method using the base scores: self-consistency, P(True), and verbalized confidence. We bold the better-performing method for each pairing and use * to denote the best-performing method across all methods.

model rate its confidence numerically (e.g., output an integer between 1-5, 1-10, 1-100, etc.). However, we found these base scores to be somewhat unreliable. Instead, we ask the LLM to rate its confidence in each individual claim using integers between 1 and 5. We then output a weighted sum of the next-token probabilities for the tokens "1" through "5": $\sum_{r=1}^{5} r \times P(r)$.

5 EMPIRICAL RESULTS

To assess the efficacy of the methods introduced in Section 3, we present results for the task of biography generation using outputs from Llama 2 7B Chat [Touvron et al., 2023] and Mistral 7B Instruct v0.2 [Jiang et al., 2023]. We randomly split the entities into 80-20 calibration-test splits, averaging results over 10 randomly generated splits.

We stress that the primary goal (and novelty) of our work is to evaluate—as it pertains to metrics for uncertainty quantification—the (1) efficacy and failures of marginal methods and (2) the extent to which multi-group methods improve over them. In supplementary results found in Appendix A, we include additional analysis comparing how marginal and group-conditional methods behave. Tables 11, 12, 13, and 14 of the appendix provide examples of the types of outputs produced by both marginal and multivalid conformal methods.

Calibration. In Table 1, we report ASCE, max gASCE, and mean gASCE, comparing each calibration method (HB, PS) against its multicalibration counterpart (IGHB, GCULR) across various base scoring functions. We find that marginal calibration methods (HB, PS) are able to correct the uncalibrated uncertainty scores, significantly decreasing ASCE. However, when examining max and mean gASCE, we find that these methods do not ensure strong guarantees for uncertainty when evaluating within different subgroups. The discrepancy is particularity large in cases where the marginal method performs well w.r.t. marginal ASCE. For example, when applying PS to self-consistency scores and comparing ASCE to max group gASCE, we observe that there exists some subgroup for which the calibration error is approximately **263x** and **198x** (for Llama and Mistral output respectively) worse on that particular subgroup compared to the dataset as a whole.

In contrast, the multicalibration variants of both the patching (IGHB) and linear regression (GCULR) techniques significantly outperform HB and PS in terms of max and mean gASCE across all experimental settings. Our results provide strong evidence that regardless of the model, base scoring function, or algorithm type, incorporating information about the subgroups in some meaningful way will substantially correct biases that marginal methods exhibit.

Even more surprising is that when considering just

Model	Base Score	Metric	Uncalibrated	HB	IGHB	PS	GCULR
Llama	self-consistency	marginal	0.475	0.169	**0.148**	0.152	**0.143***
		group max	0.535	0.323	**0.247**	0.285	**0.235***
		group mean	0.479	0.169	**0.148**	0.152	**0.143***
	P(True)	marginal	0.274	0.165	**0.152**	0.157	**0.149***
		group max	0.341	0.315	**0.261**	0.305	**0.250***
		group mean	0.277	0.165	**0.152**	0.157	**0.148***
	verb. conf.	marginal	0.177	0.161	**0.152**	0.161	**0.150***
		group max	0.270	0.311	**0.253**	0.311	**0.248***
		group mean	0.177	0.160	**0.152**	0.160	**0.149***
Mistral	self-consistency	marginal	0.471	0.186	**0.164**	0.159	**0.152***
		group max	0.554	0.333	**0.285**	0.250	**0.235***
		group mean	0.477	0.186	**0.164**	0.158	**0.152***
	P(True)	marginal	0.237	0.175	**0.164**	0.174	**0.161***
		group max	0.304	0.318	**0.259**	0.317	**0.249***
		group mean	0.237	0.175	**0.164**	0.174	**0.160***
	verb. conf.	marginal	0.397	0.175	**0.164**	0.175	**0.161***
		group max	0.427	0.318	**0.259**	0.318	**0.249***
		group mean	0.398	0.175	**0.164**	0.174	**0.160***

Table 2: We generate biographies using Llama 2 7B Chat and Mistral 7B Instruct for entities from BIO-NQ and compare each calibration method (HB, PS) against its multicalibration counterpart (IGHB, GCULR) on **Brier score** (↓ better) **marginally** across the entire dataset, as well as within each subgroup (in terms of **max** and **mean** over all groups). We test each method using the base scores: self-consistency, P(True), and verbalized confidence. We bold the better-performing method for each pairing and use * to denote the best-performing method across all methods.

(marginal) ASCE across the entire dataset, incorporating group features improves performance as well. Aside from the experiments calibrating verbalized confidence scores on Llama 2 7B Chat generations, where HB and PS perform particularly well, the multicalibration variant outperforms the marginal method every time, with GCULR being the best method in almost all cases. While improving marginal calibration is not the primary focus of our work, these results suggest that even if one does not specifically require parity for specific subgroups, collecting additional group features and applying multicalibration (as opposed to vanilla calibration) can still be extremely beneficial for generating better-calibrated uncertainty scores.

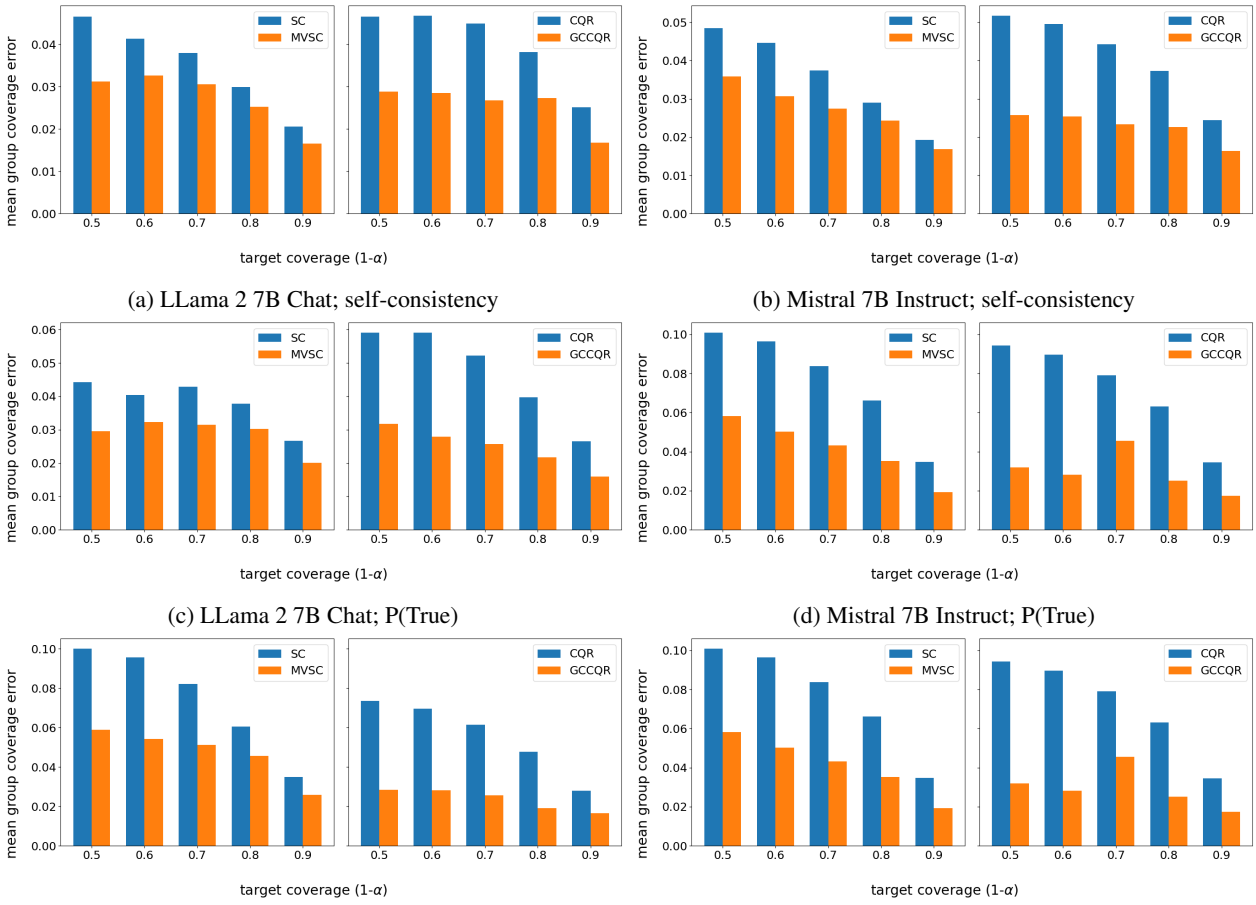
Finally, to evaluate fact-level uncertainty more holistically, we also consider the Brier score, which is the mean squared error between the uncertainty score function $f(X)$ and the true label Y. While not a direct measure of (multi)calibration like ASCE, the Brier score is still useful in certain settings for quantifying the efficacy of the algorithms we consider, quantifying desirable properties of calibration that are not captured by calibration error [Bröcker, 2009, Liu et al., 2025]. In Table 2, we report the Brier score, also both marginally and across groups. Similar to our analysis of ASCE, we again see that standard calibration methods exhibit failures when evaluating at the group level. However, IGHB and GCULR outperform HB and PS respectively across all metrics.

Conformal prediction. For the problem of uncertainty at the biography level, we apply the vanilla conformal prediction methods SC and CQR and their multivalid counterparts, MVSC and GCCQR.[6] We choose target coverages of between 0.5 to 0.9, evaluating on biographies generated by Llama 2 7B Chat and Mistral 7B Instruct.[7]

We corroborate Mohri and Hashimoto [2024]'s findings that (standard) conformal prediction methods are able to achieve close to perfect coverage on biography generation. Specifically, we show that both SC and CQR achieve target coverages (Appendix A, Figure 2). Moreover, there is little difference between the two in terms of the average number of abstentions and facts per biography retained. However, when evaluating coverage across individual subgroups, we find that both methods have some level of error. In Figure 1, we compare the mean absolute coverage error across all subgroups for each target coverage and find that SC

[6]To compare methods qualitatively, we provide illustrative example outputs in Appendix A, Tables 11, 12, 13, and 14.

[7]Although we evaluate on a wide set of target coverages $1-\alpha$, conformal prediction makes more sense only for higher target coverages (e.g., 0.8 or higher), since lower coverage guarantees can often be too weak to be useful in practice.



(a) LLama 2 7B Chat; self-consistency

(b) Mistral 7B Instruct; self-consistency

(c) LLama 2 7B Chat; P(True)

(d) Mistral 7B Instruct; P(True)

(e) LLama 2 7B Chat; verbalized confidence

(f) Mistral 7B Instruct; verbalized confidence

Figure 1: For each target coverage, we run conformal methods (**blue**: SC, CQR) and their multigroup counterparts (**orange**: MVSC, GCCQR) on BIO-NQ using the following base uncertainty scoring functions: **(a, b)** self-consistency, **(c, d)** P(True), and **(e, f)** verbalized confidence. We evaluate on generations from **(a, c, e)** Llama 2 7B Chat and **(b, d, f)** Mistral 7B Instruct. We calculate the average coverage error across all groups and plot them side by side for each pairing.

and CQR exhibit high mean errors (of up to 0.1 in some cases), despite achieving almost no error when evaluated (marginally) across the entire dataset (Figure 2).

Again, we investigate whether incorporating subgroup information can correct these biases. Here, the message is clear—multivalid conformal methods improve coverage error at the group level, regardless of the model, base scoring function, or algorithm type (Figure 1). We note however that we do not observe the same performance gains as found for calibration (Table 1), where group-conditional methods sometimes outperform marginal ones by an entire order of magnitude. This finding may result in part due to the smaller calibration set or the possibility that (multivalid) conformal prediction for LLMs is a more challenging problem. We leave further investigation of this observation to future work.

6 CONCLUSION

In this paper, we conduct an extensive study on uncertainty quantification for long-form text generation. We focus on two forms of uncertainty—claim-level (calibration) and biography-level (conformal prediction)—and present a variety of methods for these settings. We empirically validate that marginal methods for calibration and conformal prediction perform well when evaluated across the entire dataset. However, when looking at subgroup performance, we find that performance consistently degrades. Introducing two categories of algorithms (iterative patching and linear regression), we demonstrate that by accounting for additional groups, multicalibration and multivalid conformal prediction methods correct the aforementioned biases of marginal-guarantee counterparts. We consider these empirical results to establish a benchmark for this setting and hope that our findings will motivate future work in this area.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *arXiv preprint arXiv:2410.22257*, 2024.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *arXiv preprint arXiv:2406.09714*, 2024.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. *arXiv preprint arXiv:2404.04689*, 2024.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, 2023.

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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *CoRR*, 2022.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 2023.

Terrance Liu, Shuyi Wang, Daniel Preotiuc-Pietro, Yash Chandarana, and Chirag Gupta. Calibrating llms for text-to-sql parsing by leveraging sub-clause frequencies. *arXiv preprint arXiv:2505.23804*, 2025.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023a.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. Nonparametric masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2097–2118, 2023b.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 1999.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. *Unpublished Lecture Notes*, 2022.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Anthony Sicilia, Hyunwoo Kim, Khyathi Raghavi Chandu, Malihe Alikhani, and Jack Hessel. Deal, or no deal (or who knows)? forecasting uncertainty in conversations using large language models. *arXiv preprint arXiv:2402.03284*, 2024.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Lyudmila Rvanova, Sergey Petrakov, Alexander Panchenko, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *CoRR*, 2024.

Ante Wang, Linfeng Song, Baolin Peng, Ye Tian, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Fine-grained self-endorsement improves factuality and reasoning. *arXiv preprint arXiv:2402.15631*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, 2023.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, 2023.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*, 2023.

Multi-group Uncertainty Quantification for Long-form Text Generation (Supplementary Material)

Terrance Liu[1] **Zhiwei Steven Wu**[1]

[1]Carnegie Mellon University

A ADDITIONAL ANALYSIS AND RESULTS

We provide additional analysis of the methods studied in our paper. For conciseness, we conduct this analysis on methods applied to *self-consistency* scores only. We note, however, that similar findings can be made when applying such methods to P(True) or verbalized confidence.

		HB	IGHB	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **nationality** = *EU/ME*	0.0771	0.0148	-0.0623
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0540	0.0055	-0.0485
	# Wiki prop. = *Low* & **sport** = *False*	0.0429	0.0031	-0.0398
	# Wiki prop. = *Low* & **sex or gender** = *female*	0.0395	0.0033	-0.0362
	# Wiki prop. = *Low* & **nationality** = *NA*	0.0379	0.0042	-0.0337
Min Δ	**# Wiki prop.** = *Medium* & **nationality** = *APAC*	0.0114	0.0088	-0.0026

Table 3: [**Calibration** on *self-consistency* scores] Using outputs from **Llama 2 7B Chat** on BIO-NQ, we calculate the ASCE for each group using **HB** and **IGHB** as well as the difference in ASCE (Δ) between the two methods. First, we then present the top 5 groups according (Δ) where **top** corresponds to groups for which the multicalibration method achieves the biggest improvement (most negative change Δ). In our experiments, we find that IGHB improves over HB for all groups, and so as reference, we also present the group with the minimum amount of change between IGHB and HB.

		HB	IGHB	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **nationality** = *EU/ME*	0.0837	0.0166	-0.0671
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0718	0.0097	-0.0621
	# Wiki prop. = *Low* & **sport** = *False*	0.0505	0.0059	-0.0446
	# Wiki prop. = *Low* & **nationality** = *NA*	0.0458	0.0088	-0.0370
	# Wiki prop. = *Low* & **sex or gender** = *female*	0.0414	0.0045	-0.0369
Min Δ	**# Wiki prop.** = *Medium* & **nationality** = *APAC*	0.0118	0.0098	-0.0021

Table 4: [**Calibration** on *self-consistency* scores] Using outputs from **Mistral 7B Instruct** on BIO-NQ, we calculate the ASCE for each group using **HB** and **IGHB** as well as the difference in ASCE (Δ) between the two methods. First, we then present the top 5 groups according (Δ) where **top** corresponds to groups for which the multicalibration method achieves the biggest improvement (most negative change Δ). In our experiments, we find that IGHB improves over HB for all groups, and so as reference, we also present the group with the minimum amount of change between IGHB and HB.

		PS	GCULR	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **nationality** = *EU/ME*	0.0579	0.0063	-0.0516
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0375	0.0030	-0.0345
	# Wiki prop. = *Low* & **sport** = *False*	0.0266	0.0013	-0.0253
	# Wiki prop. = *Low* & **sex or gender** = *female*	0.0243	0.0034	-0.0209
	# Wiki prop. = *Low* & **nationality** = *NA*	0.0338	0.0031	-0.0306
	Mean	0.0338	0.0031	-0.0306
Bottom 5 Δ	**has IMDb ID** = *True* & **nationality** = *APAC*	0.0015	0.0018	0.0003
	# Wiki prop. = *Low* & **sport** = *True*	0.0028	0.0033	0.0005
	has IMDb ID = *False* & **nationality** = *NA*	0.0005	0.0012	0.0007
	has IMDb ID = *False* & **sex or gender** = *female*	0.0011	0.0019	0.0008
	nationality = *APAC* & **sex or gender** = *female*	0.0016	0.0024	0.0008
	Mean	0.0016	0.0024	0.0008

Table 5: [**Calibration** on *self-consistency* scores] Using outputs from **Llama 2 7B Chat** on BIO-NQ, we calculate the ASCE for each group using **PS** and **GCULR** as well as the difference in ASCE (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5. We observe that GCULR greatly improves over PS among the top 5 groups, and even in the cases where GCULR worsens ASCE compared to PS, we find that the errors are already extremely small for both PS and GCULR.

		PS	GCULR	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **nationality** = *EU/ME*	0.0573	0.0049	-0.0524
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0488	0.0035	-0.0453
	# Wiki prop. = *Low* & **sport** = *False*	0.0291	0.0011	-0.0280
	# Wiki prop. = *Low* & **nationality** = *NA*	0.0253	0.0018	-0.0235
	# Wiki prop. = *Low* & **sex or gender** = *female*	0.0226	0.0026	-0.0200
	Mean	0.0366	0.0028	-0.0338
Bottom 5 Δ	**# Wiki prop.** = *Medium* & **nationality** = *APAC*	0.0018	0.0019	0.0001
	nationality = *APAC* & **sex or gender** = *female*	0.0020	0.0027	0.0008
	has IMDb ID = *False* & **sex or gender** = *female*	0.0011	0.0022	0.0011
	has IMDb ID = *True* & **nationality** = *APAC*	0.0010	0.0022	0.0012
	# Wiki prop. = *Low* & **sport** = *True*	0.0021	0.0041	0.0020
	Mean	0.0016	0.0026	0.0011

Table 6: [**Calibration** on *self-consistency* scores] Using outputs from **Mistral 7B Instruct** on BIO-NQ, we calculate the ASCE for each group using **PS** and **GCULR** as well as the difference in ASCE (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5. We observe that GCULR greatly improves over PS among the top 5 groups, and even in the cases where GCULR worsens ASCE compared to PS, we find that the errors are already extremely small for both PS and GCULR.

Calibration. In Section 5, we demonstrate that multicalibration methods (IGHB, GCULR) significantly outperform standard calibration methods (HB, PS) with respect to calibration error both marginally and and within groups (max and mean gASCE). In this section, we further examine group calibration error, specifically looking at which groups do multicalibration methods improve over marginal methods most.

First, in Tables 3 and 4, we compare HB to IGHB for outputs from LLAMA 2 7B CHAT and MISTRAL 8B INSTRUCT, calculating the difference Δ in ASCE between the two methods for each group. Interestingly, we find that IGHB improves over HB for every group. We note that this finding is expected when one has access to the true data distribution. In our case, we implement IGHB using the calibration set (since we do not have access to the true data distribution), suggesting that the distributions for our calibration and test sets are still close enough such that IGHB is able to achieve such a strong result. Therefore, we present in Tables 3 and 4 the top 5 groups in terms of improvement Δ of IGHB compared to HB. For reference, we also present results for the group with the smallest improvement (to show the minimum improvement of the method).

Next, in Tables 5 and 6, we compare PS to GCULR for outputs from LLAMA 2 7B CHAT and MISTRAL 8B INSTRUCT. Upon initial inspection, we find that unlike for HB and IGHB, GCULR does not improve ASCE for every single group when compared to its standard variant, PS. Thus, in Tables 3 and 4, we instead show the top and bottom 5 groups in terms of improvement Δ of GCULR over PS. As shown in these results, like IGHB, GCULR is able to improve ASCE by a large margin (top 5 Δ). Moreover, we find that among groups (bottom 5 Δ) where GCULR is not able to improve ASCE, the calibration errors of PS are already very small (≤ 0.0028 for Llama 2 and ≤ 0.0021 for Mistral). While GCULR does worsen ASCE for these 5 groups, the mean difference Δ is only 0.0008 and 0.0011 for LLAMA 2 7B CHAT and MISTRAL 8B INSTRUCT, thereby achieving still small calibration errors. In comparison, when GCULR does correct ASCE for subgroups, it does so by large margin, with mean reduction in error of 0.0306 and 0.0338 respectively. Consequently, we still see large improvements for overall mean and max gASCE when comparing GCULR to PS (as shown in Table 1 of the main body).

Finally, we note that in all tables Tables 3, 4, 5, and 6, we observe the same set of groups in top and bottom 5, sorted by difference in ASCE Δ. For example, regardless of model (Llama vs Mistral) or algorithm type (iterative patching vs regression-based), the top 5 groups (and their order) are exactly the same. Similarly, we find that for both models, the group with the smallest improvement is **# Wiki prop.** = *Medium* & **nationality** = *APAC*. These observations suggest that our findings are not unique to either the model choice or calibration algorithm type.

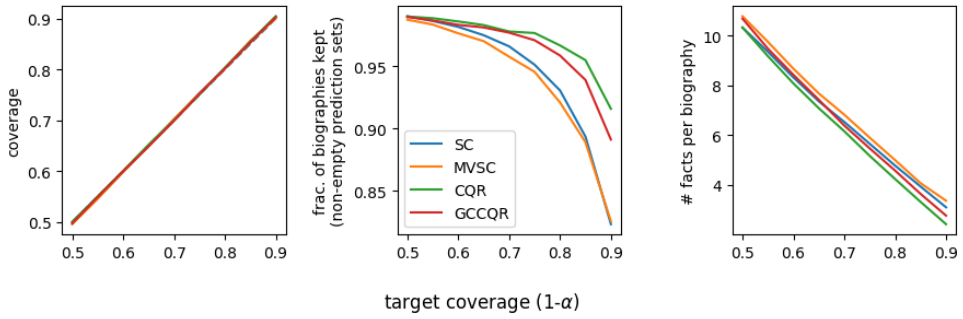
Looking specifically at which groups does multicalibration correct the most (top 5 Δ), we see that our models are most miscalibrated w.r.t. groups where the # Wikidata properties is low, suggesting that standard calibration methods (HB, PS) are miscalibrated when it comes to quantifying uncertainty for individuals whose information is not prevalent on the Internet (and therefore most likely do not appear as often in training data used to train LLMs today). Fortunately, however, incorporating group information (as is done in IGHB and GCULR) helps alleviate this issue (i.e., in Tables 3 and 4, the mean ASCE of GCULR for the top and bottom groups is fairly close).

Conformal Prediction. In Figure 2, we provide additional information about the prediction sets outputted by our various conformal methods on BIO-NQ. On the left panel, we plot the empirical coverage achieved against the target coverage. Figure 2 demonstrates that all methods achieve the target (marginal) coverage. On the **middle** panel, we plot the fraction of biographies retained (i.e., non-abstentions) for each method against the target coverage level, while on the **right** panel, we plot the number of facts per biography retained. Generally speaking, all methods retain about the same number of facts per biography. We also observe that to achieve the same target coverage, with SC and MVSC generally retaining fewer biographies (i.e., more abstentions) when compared to CQR and GCCQR. However, when comparing each conformal method (SC and CQR) to their multivalid counterparts (MVSC and GCCQR), we again observe that there are very little differences between them.

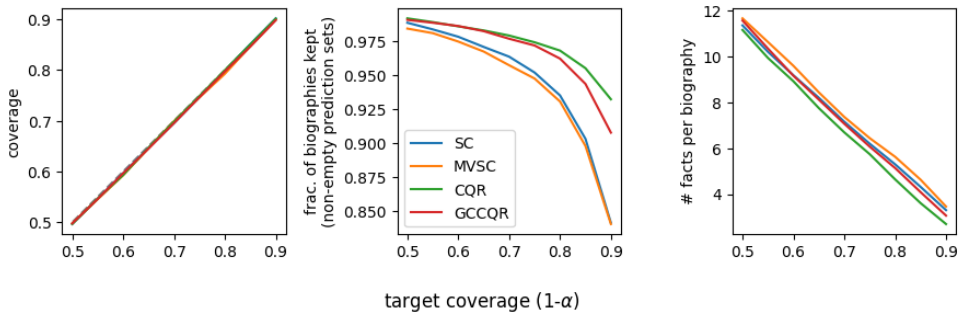
To help illustrate how different conformal methods (e.g., standard conformal vs. the multivalid counterpart) affect the final output text (i.e., subsets of retained claims), we provide examples[1] outputted by models on BIO-NQ. In Tables 11 and 12, we demonstrate how multivalid conformal methods can produce sets with additional claims retained. Moreover, in some cases, standard conformal methods (SC, CQR) may produce empty sets (abstain) while their multivalid counterparts do not (Tables 13 and 14).

Finally, like in the section above, we again examine what groups do multivalid conformal methods improve over standard

[1]Note that these examples are meant to be illustrative—measuring actual effectiveness of conformal prediction methods must done at the group or dataset level (e.g., Figure 1).



(a) LLama 2 7B Chat



(b) Mistral 7B Instruct

Figure 2: We report additional metrics for conformal predictions techniques when evaluated on biographies generated for BIO-NQ. Here, we use *self-consistency* as our base uncertainty score function. On the **top** row, we present these metrics for outputs from LLama 2 7B Chat, and on the **bottom**, Mistral 7B Instruct. On the **left** panel, we plot the empirical coverage achieved against the target coverage. On the **middle** panel, we plot the fraction of biographies retained for each method against the target coverage level. Finally, on the **right** panel, we plot the number of facts per biography retained, again against the target coverage level.

		SC	MVSC	Δ
Top 5 Δ	**# Wiki prop.** = *Very High* & **has IMDb ID** = *True*	0.0318	0.0137	-0.0182
	# Wiki prop. = *Low* & **nationality** = *NA*	0.0325	0.0153	-0.0171
	# Wiki prop. = *High* & **sex or gender** = *Female*	0.0312	0.0183	-0.0129
	has IMDb ID = *True* & **plays pro sport** = *True*	0.0427	0.0298	-0.0129
	# Wiki prop. = *Very High*	0.0256	0.0131	-0.0125
	Mean	0.0328	0.0180	-0.0147
Bottom 5 Δ	**has IMDb ID** = *False* & **nationality** = *APAC*	0.0272	0.0401	0.0129
	# Wiki prop. = *High* & **has IMDb ID** = *False*	0.0180	0.0292	0.0112
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0183	0.0216	0.0033
	# Wiki prop. = *Very High* & **has IMDb ID** = *False*	0.0259	0.0283	0.0024
	nationality = *APAC* & **plays pro sport** = *False*	0.0231	0.0244	0.0013
	Mean	0.0225	0.0287	0.0062

Table 7: [**Conformal** on *self-consistency* scores] Using outputs from **Llama 2 7B Chat** on BIO-NQ, we calculate the coverage error (for a target coverage of 90%) for each group using **SC** and **MVSC** as well as the difference in coverage error (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5.

		SC	MVSC	Δ
Top 5 Δ	**# Wiki prop.** = *Very High* & **nationality** = *EU/ME*	0.0423	0.0235	-0.0188
	# Wiki prop. = *Low* & **has IMDb ID** = *True*	0.0282	0.0112	-0.0170
	# Wiki prop. = *Very High* & **sex or gender** = *Male*	0.0335	0.0190	-0.0145
	has IMDb ID = *Medium* & **plays pro sport** = *True*	0.0404	0.0269	-0.0135
	# Wiki prop. = *Very High* & **has IMDb ID** = *False*	0.0371	0.0240	-0.0131
	Mean	0.0363	0.0209	-0.0154
Bottom 5 Δ	**nationality** = *NA* & **plays pro sport** = *True*	0.0173	0.0282	0.0109
	has IMDb ID = *False* & **plays pro sport** = *True*	0.0221	0.0279	0.0058
	plays pro sport = *False* & **sex or gender** = *Female*	0.0138	0.0192	0.0055
	sex or gender = *Female*	0.0143	0.0193	0.0050
	has IMDb ID = *True* & **sex or gender** = *Female*	0.0145	0.0190	0.0044
	Mean	0.0164	0.0227	0.0063

Table 8: [**Conformal** on *self-consistency* scores] Using outputs from **Mistral 7B Instruct** on BIO-NQ, we calculate the coverage error (for a target coverage of 90%) for each group using **SC** and **MVSC** as well as the difference in coverage error (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5.

		CQR	GCCQR	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **sport** = *False*	0.0652	0.0190	-0.0463
	# Wiki prop. = *Low* & **IMDb ID** = *True*	0.0564	0.0153	-0.0411
	# Wiki prop. = *Low*	0.0556	0.0167	-0.0389
	# Wiki prop. = *Low* & **IMDb ID** = *False*	0.0565	0.0226	-0.0339
	# Wiki prop. = *Low* & **sex or gender** = *Male*	0.0556	0.0222	-0.0334
	Mean	0.0579	0.0192	-0.0387
Bottom 5 Δ	**nationality** = *APAC* & **sport** = *False*	0.0150	0.0294	0.0143
	# Wiki prop. = *Very High* & **IMDb ID** = *False*	0.0193	0.0335	0.0142
	nationality = *APAC* & **sex or gender** = *Male*	0.0151	0.0227	0.0076
	# Wiki prop. = *Medium* & **nationality** = *EU/ME*	0.0196	0.0268	0.0072
	# Wiki prop. = *Medium* & **sex or gender** = *Female*	0.0183	0.0241	0.0058
	Mean	0.0175	0.0273	0.0098

Table 9: [**Conformal** on *self-consistency* scores] Using outputs from **Llama 2 7B Chat** on BIO-NQ, we calculate the coverage error (for a target coverage of 90%) for each group using **CQR** and **GCCQR** as well as the difference in coverage error (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5.

		CQR	GCCQR	Δ
Top 5 Δ	**# Wiki prop.** = *Low* & **nationality** = *NA*	0.0795	0.0135	-0.0661
	# Wiki prop. = *Low* & **IMDb ID** = *True*	0.0758	0.0173	-0.0585
	# Wiki prop. = *Low* & **sport** = *False*	0.0746	0.0175	-0.0571
	# Wiki prop. = *Low*	0.0662	0.0121	-0.0541
	# Wiki prop. = *Low* & **sex or gender** = *Male*	0.0698	0.0244	-0.0454
	Mean	0.0732	0.0170	-0.0562
Bottom 5 Δ	**IMDb ID** = *False* & **nationality** = *EU/ME*	0.0193	0.0331	0.0138
	nationality = *NA* & **sport** = *True*	0.0155	0.0283	0.0128
	IMDb ID = *False* & **sport** = *False*	0.0114	0.0194	0.0080
	nationality = *EU/ME* & **sex or gender** = *Female*	0.0275	0.0352	0.0077
	IMDb ID = *False* & **sport** = *True*	0.0267	0.0325	0.0057
	Mean	0.0201	0.0297	0.0096

Table 10: [**Conformal** on *self-consistency* scores] Using outputs from **Mistral 7B Instruct** on BIO-NQ, we calculate the coverage error (for a target coverage of 90%) for each group using **CQR** and **GCCQR** as well as the difference in coverage error (Δ) between the two methods. We then present the top and bottom 5 groups according (Δ) where **top** corresponds to groups for which the multivalid method achieves the biggest improvement (most negative change Δ). In addition, we calculate the mean values for the top and bottom 5.

methods on the most, where in this case, we instead calculate the difference Δ in coverage error (at target coverage of 90%) between each pairing of conformal and multivalid conformal methods. In particular, we display the top and bottom 5 groups in terms of difference Δ in Tables 7, 8, 9, and 10

Our findings show that for the topic of factuality in long-form text generation, multivalid conformal prediction is a more challenging problem when compared to calibration. As shown in Figure 1, multivalid methods (GCCQR and MVSC) consistently (at all target coverages) outperform standard conformal methods (SC and CQR) w.r.t. group coverage error. Tables 7, 8, 9, and 10 corroborate this finding, showing that the mean coverage difference Δ for the top groups is larger (at a minimum, 2.41x more for MVSC and 3.95x more for GCCQR), demonstrating that multivalid methods tend to improve coverage error on groups more than it worsens it (for other groups), thereby achieving a better mean group coverage error overall. However, the improvements are not as stark as those found in Tables 3, 4, 5, and 6 for calibration error, suggesting that multivalid conformal prediction may be a harder problem overall.

When looking at which groups do multivalid conformal methods improve the most on, we find no consistent patterns. However, we do observe that all groups for which MVSC or GCCQR improve the most on are related to the number of Wikidata properties. Interesting, we do observe that CQR does quite poorly on groups containing people with a low number of Wikidata properties, mirroring our findings for calibration above. Like in multicalibration, GCCQR is able to significanly improve coverage error for these groups. Lastly, we note that CQR seems to achieve worse group coverage than that of SC. For example, on outputs from MISTRAL 7B INSTRUCT, the mean coverage error among the top 5 groups is 0.0732 for CQR compared to 0.0363 for SC. However, we find that for both models (Tables 9, and 10), GCCQR is able to still reduce coverage errors to levels similar to that of MVSC (Tables 7, and 8).

Claims	SC	MVSC	CQR	GCCQR
Henry Cavill was born in Jersey, Channel Islands.	X	X	X	X
Henry Cavill has reprised the role of Superman in "Batman v Superman: Dawn of Justice" (2016).		X	X	X
Henry Cavill gained international recognition for his portrayal of Superman in the DC Extended Universe.		X		X
Henry Cavill has reprised the role of Superman in "Justice League" (2017).		X		X
Henry Cavill is British.				X
Henry Cavill is also known for his philanthropic work.				X
Henry Cavill is an actor.				X
Henry Cavill was born on May 5th, 1983.				X
Henry Cavill's performance in the role of Superman has been widely praised.				X
Amy Winehouse left a lasting impact on the music industry.	X	X	X	X
Amy Winehouse released her follow-up album, "Back to Black," in 2006.	X	X	X	X
Amy Winehouse's debut album "Frank" was released in 2003.	X	X	X	X
Amy Winehouse was a unique artist.	X	X		X
Amy Winehouse's lyrics were poignant.	X	X		X
The hit single "Rehab" contributed to the album's success.	X	X		X
Winehouse began singing and writing songs at a young age.	X	X		X
"Frank" received critical acclaim.		X		
Amy Winehouse was a British singer and songwriter.		X		
Amy Winehouse was a talented singer-songwriter.		X		
Fans mourned the loss of Amy Winehouse.		X		
Winehouse grew up in a family of Jewish descent.		X		

Table 11: Using outputs from **Llama 2 7B Chat** on BIO-NQ, we present examples in which all conformal methods using *self-consistency* scores (at 90% target coverage) produce a subset of claims that are entirely correct. In these examples, multivalid methods (MVSC, GCCQR) retain more claims.

Claims	SC	MVSC	CQR	GCCQR
H. G. Wells is considered a pioneer of science fiction.	X	X	X	X
H.G. Wells is best remembered for H. G. Wells's works in the science fiction genre.	X	X	X	X
H.G. Wells is most famous for H. G. Wells's science fiction.	X	X	X	X
H.G. Wells was a pioneer of the science fiction genre.	X	X	X	X
H.G. Wells was a prolific writer.	X	X	X	X
H.G. Wells was born in Bromley, England.	X	X	X	X
H.G. Wells was known for H. G. Wells's science fiction works.	X	X	X	X
H.G. Wells' most famous works include "The Time Machine," "The War of the Worlds," and "The Invisible Man."	X	X	X	X
H.G. Wells was a renowned writer.	X	X		X
H.G. Wells was an English writer.		X		X
H.G. Wells died in 1946.		X		
H.G. Wells was a prolific writer, publishing over 50 books.		X		
H.G. Wells was born on September 21, 1866, in Bromley, Kent, England.		X		
H.G. Wells wrote works in various other genres, including fiction.		X		
H.G. Wells wrote works in various other genres, including social commentary.		X		
H.G. Wells' works often explored the social and political implications of scientific and technological advancements.		X		
Wells' works, such as "The Time Machine," "The War of the Worlds," and "The Invisible Man," have had a significant impact on the development of the science fiction literary genre.		X		
Heisenberg made significant contributions to quantum mechanics.	X	X	X	X
Werner Heisenberg passed away on February 1, 1976.	X	X	X	X
Werner Heisenberg studied under Arnold Sommerfeld at the University of Munich.	X	X	X	X
Werner Heisenberg was a key figure in the development of quantum mechanics.	X	X	X	X
Werner Heisenberg was born in Wurzberg, Germany in 1901.	X	X	X	X
Werner Heisenberg's work had a profound impact on the field of physics.	X	X	X	X
Heisenberg played a pioneering role in quantum theory.		X		X
Werner Heisenberg attended the University of Munich.		X		X
Werner Heisenberg is best known for his uncertainty principle.		X		X
Werner Heisenberg's contributions paved the way for the development of quantum mechanics.		X		X
Werner Heisenberg's work revolutionized the field of physics.		X		X

Table 12: Using outputs from **Mistral 7B Instruct** on BIO-NQ, we present examples in which all conformal methods using *self-consistency* (at 90% target coverage) produce a subset of claims that are entirely correct. In these examples, multivalid methods (MVSC, GCCQR) retain more claims.

	SC	MVSC	CQR	GCCQR
Anaximander believed that the universe is infinite.		X		X
Anaximander came from a noble family.		X		X
Anaximander was born in Miletus, a city in the ancient Greek world.		X		X
Anaximander's work has survived to the present day.		X		X
Despite his contributions to philosophy, Anaximander's life remains somewhat shrouded in mystery.		X		X
Merritt Butrick is best known for his roles in the Star Trek franchise.		X		X
Merritt Butrick was an American actor.		X		X
Merritt Butrick contributed to the Star Trek franchise.				X

Table 13: Using outputs from **Llama 2 7B Chat** on BIO-NQ, we present examples in which all conformal methods (at 90% target coverage) produce a subset of claims that are entirely correct. In these examples, the multivalid methods (MVSC, GCCQR) output nonempty steps while the standard conformal methods (SC, CQR) do not.

	SC	MVSC	CQR	GCCQR
Bessel van der Kolk has written extensively on the connection between the brain, mind, and body in the healing of trauma.		X	X	X
Bessel van der Kolk is a world-renowned Dutch-American psychiatrist.		X	X	X
Bessel van der Kolk's work has had a significant impact on the understanding and treatment of trauma.		X		X
Richard Chamberlain continues to act.				X
Richard Chamberlain has received numerous awards and accolades throughout his career.				X
Richard Chamberlain was born on March 31, 1934, in Beverly Hills, California.				X

Table 14: Using outputs from **Mistral 7B Instruct** on BIO-NQ, we present examples in which all conformal methods using *self-consistency* (at 90% target coverage) produce a subset of claims that are entirely correct. In these examples, we have that either SC or CQR produce empty sets while their multivalid counterparts (MVSC and GCCQR respectively) do not.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 BIOGRAPHY GENERATION AND FACTUALITY EVALUATION

While the ground truth score must be human-annotated, Min et al. [2023a] show that FACTSCORE can be approximated by an automated process that leverages an LLM (i.e., ChatGPT and LLaMa-7B) and natural language retrieval. Following Min et al. [2023a], we also use an LLM to automate the annotation process. For some input person, **[ENTITY]**, we prompt a large language model with the following:

[INST] Question: Tell me a bio of [ENTITY]. [/INST]

We then decompose each long-form generation into a set of atomic facts, which are then checked against some set of Wikipedia articles about the **[ENTITY]** to evaluate overall performance of language model in terms of factuality. Min et al. [2023a] demonstrate that while the evaluation process should ideally be conducted by human annotators, using large language models (i.e., ChatGPT and LLama 1 7B) to both decompose long-form generations and check against Wikipedia articles serves as a very good proxy for human annotation.

Following this general framework for automated evaluation, we use Llama 2 7B Chat to decompose each generation **[GEN_BIO]** with the following prompt:

[INST] «SYS» Break down the following input into a set of small, independent claims. You must not add additional information. Output the claims as a numbered list separated by a new line. The subject of each line should be [ENTITY]. «/SYS» Input: [GEN_BIO] [/INST]

For checking each atomic fact against Wikpedia, we directly use the code released by Min et al. [2023a], which first conducts passage retrieval via Generalizable T5-based Retrievers [Liu et al., 2023] to find relevant articles from a dump of Wikipedia (dated 2023-04-01) and then prompts an LLM (i.e., ChatGPT or Llama 1 7B) to predict whether each fact is supported by the retrieved passages. For our evaluation, we again use Llama 2 7B Chat. Finally, these predictions are ensembled with predictions using likelihood estimates derived from a nonparametric masked language model [Min et al., 2023b].

We note that for prompting the LLMs described above, we use Hugging Face's transformer's library and generate responses with temperature set to 1.0.

B.2 BASE SCORING FUNCTIONS

Self-consistency. Instead of manually annotating which claims are contained in the additional generations, we automate the process. Specifically, we use a procedure similar to frequency scoring algorithm proposed by Wang et al. [2024] in which **(1)** a set of K most relevant claims from a reference generation is retrieved using a vanilla BM25 algorithm (to reduce computational costs). Then **(2)** an LLM is tasked to evaluate whether the target claim is supported by the set of K reference claims. In our work, we replace LLM prompting in step **(2)** with ALIGNSCORE-LARGE [Zha et al., 2023], which runs significantly faster and is reported by Zha et al. [2023] to compare favorably to LLM-based alignment methods.

P(True). We use the following prompt:

[INST] «SYS» Answer the question based on your knowledge of the topic, [TOPIC]. If you are unsure about the question, output False. «/SYS» Question: Is the following statement True or False? [CLAIM] [/INST]

Verbalized confidence. We use the following prompt:

[INST] «SYS» Given a [TOPIC]: [CLAIM] pair as input, use your knowledge about [TOPIC] to rate (on an integer scale between 1 and 5) how confident you are that the input [CLAIM] is true. «/SYS» [TOPIC]: [CLAIM] [/INST]

B.3 MISCELLANEOUS DETAILS

Datasets. Table 15 reports additional information about datasets BIO-NQ and BIO-FACTSCORE, including the number of entities and claims per biography outputted by each model.

dataset	model	# entities	total # claims	avg. # claims per bio.
BIO-NQ	Llama 2 7B Chat	8,541	206,620	24.19
	Mistral 7B Instruct	8,541	297,714	34.86
BIO-FACTSCORE	Llama 2 7B Chat	683	17,605	25.78
	Mistral 7B Instruct	683	25,283	37.02

Table 15: Statistics describing our two datasets and how many claims are generated by each LLM.

Group Attributes. For our experiments, we use the following group attributes:

- **# Wikidata properties**: For each entity, we count the number of Wikidata properties and discretize them into the following buckets: $[0, 25), [25, 50), [50 - 100), [100, \infty)$. This group serves as proxy for the amount of information available online for some given entity.
- **nationality**: Following Min et al. [2023a], who use nationality derived from Wikidata to sample their dataset of human entities, we take the property *country of citizenship* (or *place of birth* when not available) and categorize the corresponding value into the following categories defined by Min et al. [2023a]: Asia/Pacific, Europe/Middle East, North America, Latin/South America/Africa.
- **sex or gender**: We take directly the value for the Wikidata property, *sex or gender*.
- **plays professional sports**: We check whether the Wikidata entry has the property, *sport*.
- **has IMDb entry**: We check whether the Wikidata entry has the property, *IMDb ID*, to use as a proxy for whether a person has been involved in films or television series.

In total, we have $|\mathcal{G}| = 77$ subgroups. To prevent extremely uncommon groups that may exist in the Wikidata database from biasing our results, we exclude groups of size $< 5\%$ of the total test set size. Note that while we create groups using 1 and 2-way combinations for evaluation, we train the quantile regression models in CQR and GCCQR using only single attribute groups as features in order to reduce computation.

Hyperparameters. For our patching algorithms IGHB and MVSC, we set the max iterations $T = 100$. For training (multi)calibration, our logistic regression models are trained using default hyperparameters given my Sci-kit learn. For training CQR and GCCQR, we run 5-fold cross validation for each target coverage $1 - \alpha$ to optimize the ℓ_1-penalty term $C \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$

For ALIGNSCORE, we set $M = 4$ and $K = 5$. We found that ALIGNSCORE generally returns values close to 0 or 1, giving us self-consistency uncertainty scores around the 5 values $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. As a result, we evaluate all methods using $p = 5$ level sets.

GPU requirements. We use a NVIDIA A100 80GB GPU for all experiments. For obtaining results on all entities across BIO-NQ and BIO-FACTSCORE, our experiments, per LLM require approximately the following:

- Generating biographies (+ 4 additional generations for getting frequency scores): 15 hours (x5)
- Splitting atomic facts (+ 4 additional generations for getting frequency scores): 30 hours (x5)
- Checking facts against Wikipedia: 75 hours (x1)
- Calculating frequency scores via AlignScore: 10 hours (x1)

Licenses. Wikidata and Wikipedia are licensed under the Creative Commons CC0 License. Llama 2 7B is licensed under Meta's Llama 2 license. Mistral 7B Chat and Hugging Face's transformers library are licensed under Apache 2.0 license. We also make use of code released by Min et al. [2023a] under the MIT license.

C EVALUATING ON ENTITIES USED IN MIN ET. AL (2023A)

In addition to evaluating on our dataset, BIO-NQ, we construct an additional dataset using the 683 entities used by Min et al. [2023a] for their empirical evaluation. We denote this dataset as BIO-FACTSCORE and evaluate all methods using *self-consistency* as the base scoring function.

Calibration. In Table 16, we observe similar results to that on BIO-NQ—namely, multicalibrated counterparts (IGHB and GCULR) perform better than their base counterpart (HB and PS). However, we note that for Mistral 7B Instruct, PS performs the best when looking at marginal ASCE. We hypothesize that the smaller gap in ASCE between PS and GCULR may be due to the smaller training size of BIO-FACTSCORE (25,283 claims), which is roughly 10x smaller than that of BIO-NQ (297,714 claims). Lastly, with respect to Brier score, multicalibration still dominates across all metrics (Table 17).

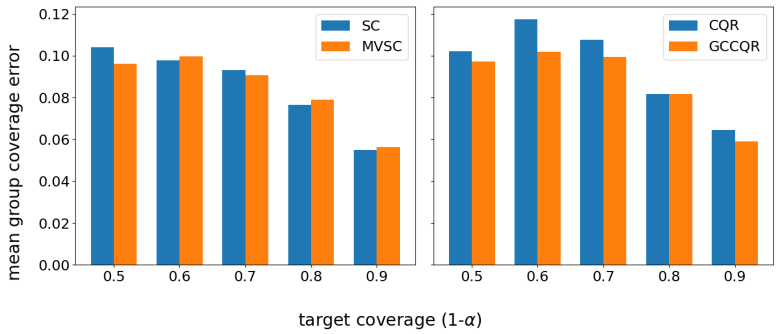
Conformal Prediction. For BIO-FACTSCORE, we observe that multivalid conformal methods *do not* improve performance across subgroups. In Figure 3, we observe very little difference in mean coverage error across groups. We hypothesize, however, that this negative result again is due to the smaller dataset size. In this case, our number of examples is the number of biographies in the dataset (683), giving us a calibration set size of 546 and test set size of 137. Further dividing the calibration and test sets into subgroups, it is possible there could simply not be enough examples per group for the distribution on the calibration set to generalize to the test set. Comparing the left panels of Figures 4 to 2, we also find that even when looking at marginal coverage, all methods perform worse (the lines deviate more from $y = x$), likely due again to the small calibration and test size.

Model	Metric	Uncalibrated	HB	IGHB	PS	GCULR
Llama 2 7B Chat	marginal	0.26830	0.00951	**0.00229**	0.00164	**0.00125***
	group max	0.48594	0.07208	**0.04088**	0.05017	**0.03519***
	group mean	0.29983	0.02848	**0.01108**	0.01659	**0.00858***
Mistral 7B Instruct	marginal	0.25496	0.01032	**0.00268**	**0.00093***	0.00146
	group max.	0.54701	0.08436	**0.04585***	0.07043	**0.04931**
	group mean	0.29435	0.03226	**0.01143**	0.01848	**0.00911***

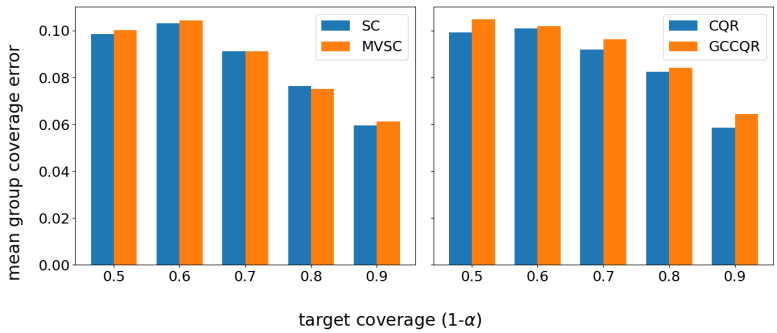
Table 16: We generate biographies for entities from BIO-FACTSCORE and compare each calibration method (HB, PS) against its multicalibration counterpart (IGHB, GCULR) on **ASCE**, **max gASCE**, and **average gASCE** (↓ better). We bold the better-performing method for each pairing. * denotes the best-performing method across all methods evaluated. All methods use *self-consistency* as their base scoring function.

Model	Metric	Uncalibrated	HB	IGHB	PS	GCULR
Llama 2 7B Chat	marginal	0.475	0.169	**0.148**	0.152	**0.143***
	group max	0.535	0.323	**0.247**	0.285	**0.235***
	group mean	0.479	0.169	**0.148**	0.152	**0.142***
Mistral 7B Instruct	marginal	0.471	0.186	**0.159**	0.164	**0.152***
	group max	0.554	0.333	**0.250**	0.285	**0.235***
	group mean	0.477	0.186	**0.158**	0.164	**0.152***

Table 17: We generate biographies for entities from BIO-FACTSCORE and compare each calibration method (HB, PS) against its multicalibration counterpart (IGHB, GCULR) on **Brier score** (↓ better) **marginally** across the entire dataset, as well as within each subgroup (in terms of **max** and **mean** over all groups). We bold the better-performing method for each pairing. * denotes the best-performing method across all methods evaluated. All methods use *self-consistency* as their base scoring function.

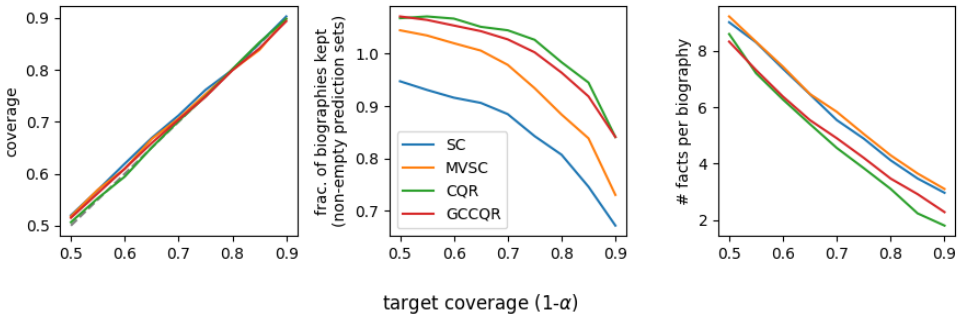


(a) LLama 2 7B Chat

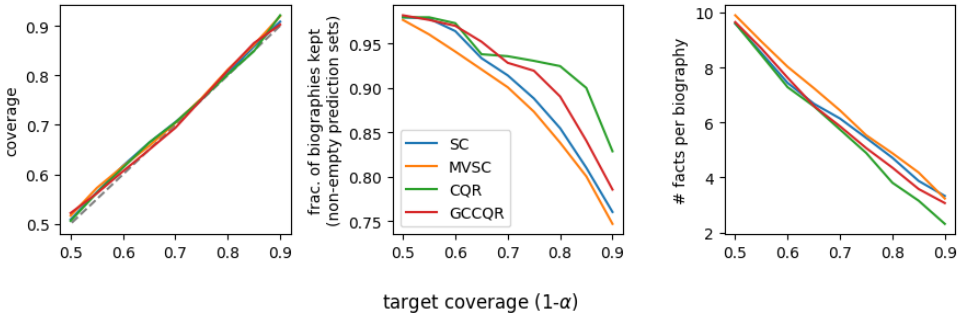


(b) Mistral 7B Instruct

Figure 3: For each target coverage, we run conformal methods (SC, CQR) and their multigroup counterparts (MVSC, GCCQR) on BIO-FACTSCORE. We evaluate on generations by (a) Llama 2 7B Chat and (b) Mistral 7B Instruct. We calculate the average coverage error across all groups and plot them side by side for each pairing. All methods use *self-consistency* as their base scoring function.



(a) LLama 2 7B Chat



(b) Mistral 7B Instruct

Figure 4: We report additional metrics for conformal predictions techniques using *self-consistency* scores when evaluated on biographies generated for BIO-FACTSCORE: On the **top** row, we present these metrics for outputs from LLama 2 7B Chat, and on the **bottom**, Mistral 7B Instruct. On the **left** panel, we plot the empirical coverage achieved against the target coverage. On the **middle** panel, we plot the fraction of biographies retained for each method against the target coverage level. Finally, on the **right** panel, we plot the number of facts per biography retained, again against the target coverage level.