# Multicalibration for Confidence Scoring in LLMs

**Gianluca Detommaso** [1] **Martin Bertran** [*1] **Riccardo Fogliato** [*1] **Aaron Roth** [1 2]

## Abstract

This paper proposes the use of "multicalibration" to yield interpretable and reliable confidence scores for outputs generated by large language models (LLMs). Multicalibration asks for calibration not just marginally, but simultaneously across various intersecting groupings of the data. We show how to form groupings for prompt/completion pairs that are correlated with the probability of correctness via two techniques: clustering within an embedding space, and "self-annotation" — querying the LLM by asking it various yes-or-no questions about the prompt. We also develop novel variants of multicalibration algorithms that offer performance improvements by reducing their tendency to overfit. Through systematic benchmarking across various question answering datasets and LLMs, we show how our techniques can yield confidence scores that provide substantial improvements in fine-grained measures of both calibration and accuracy compared to existing methods.

## 1. Introduction

Large language models (LLMs) have revolutionized text generation, with applications ranging from code development (Chen et al., 2021) to information retrieval (Zhu et al., 2023). However, alongside their impressive capabilities, LLMs possess a troubling tendency to fabricate information, generating outputs that diverge from factual reality – a phenomenon dubbed "hallucination" (Huang et al., 2023a). These hallucinations pose significant challenges to the trustworthiness and ethical deployment of LLMs, demanding the development of robust detection and mitigation strategies.

In this paper, we leverage recent "multicalibration" techniques (Hébert-Johnson et al., 2018) to produce calibrated probabilities indicating whether a generated response constitutes a hallucination. Unlike conventional calibration methods, multicalibrated probabilities are self-consistent not just marginally (i.e. on average over all examples), but also *conditionally* on various properties of the instance, which allows them to serve as more refined risk measures. Producing "risk scores" for hallucinations can provide an interpretable measure of risk which can be exposed to the user (e.g. through a coloring scheme, as in Figure 1) to communicate the risk associated with the generated content. Moreover, when those risk scores are calibrated, they are not only interpretable but "trustworthy" in the sense that they can be safely used as if they were true probabilities (Noarov et al., 2023).

| Question | Answer |
|---|---|
| The limbic system plays an important role in | regulating emotion. |
| According to aristotle, happiness is: | activity of the soul in accordance with virtue. |
| When we speak of time dilation, we mean that time | is related to space. |
| What is the biggest cause of death in children under five years old (as of 2017)? | Diarrheal diseases |

*Figure 1.* An application of multicalibration to question answering. Answers are colored from red to green according to their multicalibrated confidence scores of being a hallucination. Multicalibration is performed using Algorithm 5.

Our approach mirrors the robust assurances offered by conformal prediction, where multicalibration (of quantiles) has been used to give group-conditional guarantees (Bastani et al., 2022; Jung et al., 2022; Gibbs et al., 2023). Traditionally multicalibration has been used to give estimates of uncertainty in tabular data settings that hold conditionally on various features that are explicitly present in the data — often demographic attributes like sex or race. A key challenge in applying these techniques to hallucination detection in LLMs is a lack of such explicit features. An

[*]Equal contribution [1]AWS AI [2]University of Pennsylvania. Correspondence to: Gianluca Detommaso <detommaso.gianluca@gmail.com>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

important part of our contribution is generating features that are useful to multicalibrate with respect to — which we do both through clustering prompt embeddings, and by having the LLM itself annotate prompts with binary features via the answers to yes-or-no questions.

We note that what in many contexts, what is and is not a "hallucination" can be open to interpretation, and does not have sharp boundaries. In this study, we adopt an agnostic stance toward its definition. Specifically, we refrain from stipulating criteria for determining what constitutes "good" or "bad" generated content. Instead, we assume access to a modestly sized calibration dataset that has been annotated with binary labels. For any criterion for what constitutes a "good" vs. "bad" completion in a given context, such a dataset could be produced by human evaluators. For our work, we assume that this is given and do not take a stance on what the criterion for establishing that a given completion is "good" in a given context should be.

Our contributions are threefold: 1. We show how to apply multicalibration techniques in the context of hallucination detection in LLMs; a primary challenge here is to obtain reasonable "groups" with respect which to multicalibrate, which we do via prompt clustering and via self-annotation of prompts. 2. We introduce novel variations of multicalibration methods which yield substantial performance enhancements. 3. We systematically evaluate these techniques across diverse LLMs and question answering datasets, demonstrating their efficacy in calibration and overall performance compared to existing baselines.

**Additional Related Work** Numerous recent surveys focus on hallucinations in LLMs (Chang et al., 2023; Huang et al., 2023a; Ji et al., 2023; Rawte et al., 2023; Tonmoy et al., 2024; Zhang et al., 2023b; Guerreiro et al., 2023). The predominant focus of current research lies in binary hallucination *detection*, specifically the capacity to discern whether generated text exhibits signs of hallucination. Key contributions in this domain include (Manakul et al., 2023; Rebedea et al., 2023), which evaluate consistency, similarity, and agreement among alternative generated responses. (Kadavath et al., 2022; Friel & Sanyal, 2023) directly engage LLMs by posing inquiries about correctness or consistency within a single answer.

More closely related is a smaller body of literature that explores uncertainty quantification and confidence scoring in this context (Xiao & Wang, 2021; Verma et al., 2023; Varshney et al., 2023; Kalai & Vempala, 2023; Tian et al., 2023; Zhao et al., 2023; Chen & Mueller, 2023; Duan et al., 2023; Lin et al., 2023; Liu et al., 2023) and propose a variety of approaches to reduce hallucination generation ranging from updated beam search methods, fine-tuning, human labelling, and epistemic neural networks. Several recent papers use conformal prediction to derive sets of prompt completions, offering marginal coverage guarantees (e.g. for 90% of prompts, at least one completion in the set should be "good") (Quach et al., 2023; Kumar et al., 2023; Deutschmann et al., 2023; Ren et al., 2023; Zecchin et al., 2023). Among these, (Kumar et al., 2023) is closest to our approach but requires "group-specific" prompting strategies. The remainder focus on improving the LLM's decoding strategy and/or predictive sets, which are less suited to binary classification settings (like hallucination detection), where they are limited to $\{0\}$, $\{1\}$, and $\{0, 1\}$.

## 2. Background on (Multi)Calibration

Consider $(X, Y) \sim \mathcal{D}$ where $X \in \mathcal{X}$ indicates a prompt/completion pair, $Y$ indicates whether the completion is a hallucination given the prompt ($Y = 0$) or not ($Y = 1$), and $\mathcal{D}$ represents the joint distribution over pairs $(X, Y)$. Let $f : \mathcal{X} \mapsto [0, 1]$ denote a score representing a confidence that $Y = 1$ for the text $X$. See Section 4.1 for a discussion about possible scores $f(x)$.

Ideally, we would like to find a model $f(x)$ such that

$$f(x) = \mathbb{P}_{\mathcal{D}}(Y = 1|X = x), \quad \forall\, x \in \mathcal{X}. \tag{1}$$

However, there are two difficulties with this. First, this may not be a coherent probabilistic notion: fixing $x$, the label may be determined, and so the "probability" of $Y = 1$ may be either 0 or 1. Moreover, it is generally impossible to learn a function with this property without observing *every* possible $x$, which is impossible for extremely large sets $\mathcal{X}$, as is the case for LLM prompt/completion pairs (Lei & Wasserman, 2014). Calibration is a simple, tractable, guarantee that corresponds to a significant coarsening of the set of conditioning events in Equation 1, to the level sets of the predictor $f$.

**Definition 2.1** (Calibration)**.** Given a data distribution $\mathcal{D}$, the *bias* of a model $f$ at the $p$-th level set is defined as

$$\Delta_p(f) := \mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = p]. \tag{2}$$

Then, if $\Delta_p(f) = 0$ for all $p \in [0, 1]$ such that $\mathbb{P}_{\mathcal{D}}(f(X) = p) > 0$, we say that $f$ is *calibrated* w.r.t. $\mathcal{D}$.

We observe that Definition 2.1 can be rewritten as

$$\mathbb{P}_{\mathcal{D}}(Y = 1|f(X) = p) = p.$$

Informally, calibration is a minimal consistency condition: it states that the conditional distribution on $Y$ conditional on the prediction that $f(X) = p$ is indeed a Bernoulli distribution with bias $p$. While a perfect model satisfying (1) is calibrated, the converse is not necessarily true.

We introduce the following standard measure of calibration error (Globus-Harris et al., 2023).

**Definition 2.2** (Average squared calibration error)**.** We denote the *average squared calibration error* (ASCE) of a model $f$ w.r.t. a distribution $\mathcal{D}$ by

$$\text{ASCE}(f) := \mathbb{E}_P[\Delta_P^2(f)]. \tag{3}$$

The ASCE is computed by integrating the squared model bias over all level sets. Note that if the model $f$ is calibrated, then $\text{ASCE}(f) = 0$. The ASCE is related to the well-known expected calibration error (ECE) (Naeini et al., 2015), with the difference that the ASCE compares $Y$ against $f$, while the ECE compares accuracy against confidence (see Appendix A). ASCE is a useful measure of calibration error because it is directly related to a natural measure of accuracy: mean squared error.

**Definition 2.3** (Mean squared error)**.** We denote the *mean squared error* (MSE) of a model $f$ w.r.t. a distribution $\mathcal{D}$ by

$$\text{MSE}(f) := \mathbb{E}_{\mathcal{D}}\left[(Y - f(X))^2\right]. \tag{4}$$

The MSE is also known as *Brier score* (Brier, 1950). A bias-variance decomposition clarifies its relation to the ASCE.

**Proposition 2.4** (See, e.g. (Kohavi et al., 1996) )**.** *We have*

$$\text{MSE}(f) = \text{ASCE}(f) + \mathbb{E}_P[\text{Var}_{\mathcal{D}}(Y|f(X) = P)]. \tag{5}$$

A proof is given in Appendix B for completeness. Proposition 2.4 shows that the MSE can be decomposed as the ASCE and the variance of the data given the model, respectively measuring how calibrated the model is and how much variation in the data the model can explain. As such, the MSE is not a direct measure of calibration — when comparing two models, the model with lower squared error may still be the less well calibrated model.

### 2.1. A Simple Calibration Strategy

**Algorithm 1** Histogram Binning (HB)

1: **for all** $p \in [\frac{1}{m}]$ **do**
2: Set
$$\hat{f}(x) := \begin{cases} f'(x) + \Delta_p(f') & \text{if } x \in S_p(f'), \\ f'(x) & \text{otherwise.} \end{cases} \tag{6}$$
3: **end for**

Given a distribution $\mathcal{D}$, a model $f$, and a threshold $\alpha > 0$, our goal is to produce a new model $\hat{f}$ that is calibrated and has reduced MSE compared to $f$ — at least up to discretization error $\alpha$. Here $\alpha$ will control the number of level sets we discretize $f$ to, and hence the complexity of the model: choosing smaller values of $\alpha$ will require more data to avoid overfitting and vice versa. First, we introduce the notation for the level sets which appear as conditioning events in (2): $S_p(f) := \{f(x) = p\}$. Since it is infeasible to condition on $S_p(f)$ for all $p \in [0, 1]$, we introduce the uniform grid $[\frac{1}{m}] := \{\frac{i}{m}\}_{i=0}^m$, and define

$$f'(x) := \arg\min_{p \in [\frac{1}{m}]} |f(x) - p|, \tag{7}$$

which rounds the model $f$ to the grid $[\frac{1}{m}]$. Algorithm 1 rounds the model, then applies a constant shift in the bin $S_p(f')$, for each element $p$ in the grid. Because $\Delta_p(\hat{f}) = 0$ for all $p \in [\frac{1}{m}]$, the following result holds.

**Theorem 2.5** (See, e.g. (Roth, 2022))**.** *Algorithm 1 satisfies* $\text{ASCE}(\hat{f}) = 0$*. Furthermore, set* $B_p := \{x : |f(x) - p| \leq \frac{1}{2m}\}$*. If* $m > \frac{1}{\alpha}$*, then*

$$\text{MSE}(\hat{f}) < \text{MSE}(f) - \sum_{p \in [\frac{1}{m}]} \mathbb{P}_{\mathcal{D}}(B_p)\, \Delta_p^2(f') + \frac{\alpha^2}{4} + \alpha.$$

A proof is given in Appendix C for completeness. Theorem 2.5 shows that by replacing the level sets of the model (on a refined enough grid) with the label mean of points conditional on the level set not only guarantees calibration, but improves the MSE of the model by its initial calibration error (as measured by ASCE). This is a key property of calibration and related guarantees: enforcing it is only accuracy enhancing. Furthermore, one can prove out-of-sample generalization bounds that replace the joint distribution $\mathcal{D}$ with the empirical distribution characterized by the available data, and hence satisfy a non-asymptotic calibration guarantee — see e.g. (Roth, 2022). For simplicity, and because the generalization bounds are standard, in our exposition here we will focus on in-sample guarantees. As Algorithm 1 is closely related to histogram binning (Zadrozny & Elkan, 2001), we will refer to it with this name.

HB crucially uses the fact that the level sets $S_p(f)$ defined in Definition 2.1 are disjoint. In the following section, we will define multicalibration (Hébert-Johnson et al., 2018), a stronger calibration guarantee that imposes simultaneous requirements on non-disjoint conditioning sets.

### 2.2. Towards Multicalibration

In this section, we argue that the promise made by calibration in Definition 2.1 is too weak for the kinds of language model applications we have in mind because the performance of a model is extremely heterogeneous across different kinds of tasks that it can be used for. As an example, consider two prompt/completion pairs $x_1$ and $x_2$, respectively asking and answering a question about (1) the capital cities of US states, and (2) citations to the academic literature for theorems in functional analysis. We would expect that the probability of correctness differs substantially across these examples — and yet calibration is a *marginal* guarantee that can average over both cases. It could, for example, lead to confidence assessments that are systematically

over-confident about academic citations and systematically under-confident about state capitals. This is not in conflict with even perfect calibration. It would be better to promise that our confidence scores were calibrated conditionally on (as fine-grained information as possible about) the prompt used. These kinds of conditional calibration guarantees are what multicalibration aims for.

We now formalize the concept of *groups*. A group function $g : \mathcal{X} \to \{0, 1\}$ can be thought of as an indicator function for a group defined as a set of prompt/completion pairs: $g(x) = 1$ if $x$ is a member of the "group" and $g(x) = 0$ otherwise. The "group" induced by an indicator function $g$ is therefore $\{x \in \mathcal{X} : g(x) = 1\}$. A set of groups $\mathcal{G}$ is correspondingly identified by a set of group indicator functions. Crucially, the groups in a collection $\mathcal{G}$ can be *intersecting* — i.e. there can be multiple groups that contain the same example $x$. This corresponds to a prompt/completion pair having multiple non-mutually-exclusive attributes: for example, $x$ might simultaneously be "a question requiring high-school level knowledge" and "a question about mathematics".

#### 2.2.1. Group-Conditional Unbiasedness

**Algorithm 2** Group-Conditional Unbiased Regression

1: Set

$$\hat{f}(x) := f(x) + \sum_{g \in \mathcal{G}} \lambda_g \, g(x), \quad \text{s.t. } \{\lambda_g\}_{g \in \mathcal{G}} = \arg\min \mathrm{MSE}(\hat{f}). \tag{8}$$

Before introducing multicalibration, let us first introduce a simpler guarantee, and a simple strategy to obtain it. Calibration (Definition 2.1) requires that a model $f$ be unbiased conditional on its own level sets. Given a collection of groups $\mathcal{G}$, we can instead ask a model to be unbiased conditionally on each of the group indicator functions in $\mathcal{G}$.

**Definition 2.6** (Group-conditional unbiasedness)**.** Given a data distribution $\mathcal{D}$ and set of groups $\mathcal{G}$, we say that a model $f$ is *group-conditionally unbiased* if

$$\mathbb{E}_{\mathcal{D}}[Y - f(X) | g(X) = 1] = 0, \quad \forall\, g \in \mathcal{G}. \tag{9}$$

This condition is also known as "multi-accuracy" in the algorithmic fairness literature (Hébert-Johnson et al., 2018; Kim et al., 2019).

Since groups may be overlapping, we cannot proceed as in Section 2.1 and independently unbias the predictions within each group. Instead, we introduce Algorithm 2, initially proposed by (Gopalan et al., 2022c), in which a model $\hat{f}$ is fit by solving a linear regression problem over features defined both by the original model $f$ and the group indicator functions in $\mathcal{G}$. The following theorem is due to (Gopalan et al., 2022c). We follow the presentation of (Roth, 2022).

**Theorem 2.7.** *The model $\hat{f}$ produced in* (8) *satisfies group-conditional unbiasedness.*

A proof is given in Appendix D for completeness. Once again, with standard techniques one can prove generalization bounds for this algorithm (see (Roth, 2022)) which allows one to replace in-sample MSE with true distributional MSE; we elide these details for simplicity here. The model produced in (8) is solving a linear regression problem (minimizing squared error over a set of linear models). In Appendix D, we generalize this result to show that it holds also when we replace MSE with a cross-entropy loss — i.e. when solving logistic regression rather than linear regression. We name the latter method *Group-Conditional Unbiased Logistic Regression* (GCULR). A similar generalization is given by (Gopalan et al., 2023).

#### 2.2.2. Multicalibration

In Section 2, we defined calibration by conditioning on the level sets of the model. In Section 2.2.1, we defined group-conditional unbiasedness by conditioning on a set of groups. Multicalibration, introduced by (Hébert-Johnson et al., 2018), is a stronger guarantee that asks for unbiasedness when simultaneously conditioning on both level sets and groups. In order to rigorously define it, let us first generalize the definition of ASCE.

**Definition 2.8** (Group average squared calibration error)**.** Given a group function $g$, we denote the *group average squared calibration error* (gASCE) of a model $f$ w.r.t. a distribution $\mathcal{D}$ by

$$\mathrm{gASCE}(f, g) := \mathbb{E}_P\left[\Delta^2_{P,g}(f) | g(X) = 1\right], \tag{10}$$

where $\Delta_{p,g}(f) := \mathbb{E}_{\mathcal{D}}[Y - f(X) | S_{p,g}(f)]$ and $S_{p,g}(f) := \{f(x) = p, g(x) = 1\}$.

Unlike the ASCE, in the gASCE we do not only condition on the disjoint level sets $\{f(x) = p\}$, but also on the groups $\{g(x) = 1\}$, which allows us to quantify the calibration error independently for each group.

**Definition 2.9** (Multicalibration)**.** Given a data distribution $\mathcal{D}$ and a set of groups $\mathcal{G}$, a model $f$ is $\alpha$-approximately *multicalibrated* w.r.t. $\mathcal{D}$ and $\mathcal{G}$ if and only if

$$\mathrm{gASCE}(f, g) < \tfrac{\alpha}{\mathbb{P}_{\mathcal{D}}(g(X)=1)}, \quad \forall\, g \in \mathcal{G}. \tag{11}$$

This is an $\ell_2$ notion of multicalibration, as studied in (Globus-Harris et al., 2023). One can also study error in other metrics (e.g. $\ell_\infty$ error as in (Hébert-Johnson et al., 2018) or $\ell_1$ error as in (Gopalan et al., 2022b)).

Compared to the Definition 2.1 of calibration, multicalibration is a stronger guarantee, where standard calibration is recovered by taking $\mathcal{G} = \{\mathcal{X}\}$. Because the groups may

overlap, we cannot independently apply patches for all conditioning sets. However, one can use a similar idea with an iterative approach, which results in an algorithm that is guaranteed to satisfy multicalibration in a finite number of rounds, and decrease the MSE at every round.

**Algorithm 3** Iterative Grouped Histogram Binning (IGHB)

1: Let $m = \lceil \frac{1}{\alpha} \rceil$, $t = 0$, $f_0 := f'$.
2: **while** $\max_{g \in \mathcal{G}} \mathbb{P}_{\mathcal{D}}(g(X) = 1)\, \text{gASCE}(f, g) > \alpha$ **do**
3: Set

$$(p_t, g_t) = \underset{p \in [\frac{1}{m}],\, g \in \mathcal{G}}{\arg\max}\ \mathbb{P}_{\mathcal{D}}\left(S_{p,g}(f_t)\right)\ \Delta^2_{p,g}(f_t).$$

4: Set $h_{t+1}(x) := \begin{cases} f_t(x) + \Delta_{p_t, g_t}(f_t) & \text{if } x \in S_{p_t, g_t}(f_t), \\ f_t(x) & \text{otherwise.} \end{cases}$.
5: Set $f_{t+1} := h'_{t+1}$.
6: **end while**

IGHB (Algorithm 3) (a variant of which was first given by (Hébert-Johnson et al., 2018)) starts by checking whether $\alpha$-approximate multicalibration is satisfied. If not, it finds the conditioning event for which the gASCE is largest, and patches the model on examples that trigger that event. It iterates like this until convergence. The rounding operation makes sure that the number of level sets do not increase without bound, which guarantees that there is sufficient data to evaluate the bias on each of the conditioning events.

**Theorem 2.10.** *Algorithm 3 halts after $T < \frac{4}{\alpha^2}$ rounds and returns a model $f_T$ that is $\alpha$-approximately multicalibrated. Moreover, if the algorithm runs for $T$ rounds, then*

$$\text{MSE}(f_T) < \text{MSE}(f) - (T-1)\frac{\alpha^2}{4} + \alpha.$$

A proof can be found in (Roth, 2022) along with out-of-sample generalization guarantees. As with HB, running IGHB is only accuracy-improving.

## 3. Remedying overfitting

The multicalibration strategies outlined in 2.2.2 can build complex models (because of iterative updates on intersecting groups), and are known to be subject to overfitting in practice (see e.g. (Globus-Harris et al., 2023)). One reason for this is that the technique operates by iteratively estimating the label mean on subsets of the data defined as $\{f(x) = v, g(x) = 1\}$, which in-sample can contain very few points and thus lead to inaccurate estimates of distributional quantities. Here, we provide practical improvements to the IGHB algorithm that lead to better performance.

### 3.1. Bins as upper and lower sets

We make the following observation: rather than conditioning on the exact value of the model $f(x) = p$, we can condition on $f(x) \leq p$ (roughly speaking conditioning on values of its CDF rather than its density function), and the definition of (exact) multicalibration remains unchanged.

**Proposition 3.1.** *If $\mathbb{E}_{\mathcal{D}}[Y|f(X) = p]$ is a continuous function of $p$, the definition of perfect multicalibration (*Definition *2.9 with $\alpha = 0$) is unchanged if we replace $S_{p,g}$ by $S^{\leq}_{p,g} := \{f(x) \leq p, g(x) = 1\}$.*

A proof is in Appendix E. Proposition 3.1 suggests a definition of multicalibration that is defined on considerably larger conditioning sets. By symmetry, the same results holds for $S^{\geq}_{p,g} := \{f(x) \geq p, g(x) = 1\}$. Thus "patching" sets $S^{\geq}_{p,g}$ or $S^{\leq}_{p,g}$ within the IGHB algorithm when the model exhibits bias on them moves the model closer to multicalibration, and the same convergence analysis applies. Note that $S^{\leq}_{p,g}$ is larger for large $p$, and vice versa for $S^{\geq}_{p,g}$, hence one may want to use one or the other bin according to the value of $p$. Note that the per-bin calibration error in step 3 of Algorithm 3 is proportional to the size of the bin, implying that larger bins are likely to be patched earlier than smaller bins. Updates on large sets are less prone to overfitting because we have many samples to use to estimate their label mean. It follows that without invalidating Theorem 2.10, in order to patch the model on a sequence of considerably larger bins we can simply replace step 3 in Algorithm 3 with

$$(p_t, g_t, \tau_t) = \underset{p \in [\frac{1}{m}],\, g \in \mathcal{G}, \tau \in \{\leq, \geq\}}{\arg\max}\ \mathbb{P}_{\mathcal{D}}\left(S^{\tau}_{p,g}(f_t)\right)\ \left(\Delta^{\tau}_{p,g}(f_t)\right)^2, \tag{12}$$

where $\Delta^{\tau}_{p,g}$ is defined as $\Delta_{p,g}$ in Definition 2.2 but replacing $S_{p,g}$ with $S^{\tau}_{p,g}$. To reiterate, what makes this approach work is that (1), a model with no bias over the sets $S^{\tau}_{p,g}$ also has no bias over the sets $S_{p,g}$ and hence satisfies multicalibration, and (2) the sets $S^{\tau}_{p,g}$ are larger and hence reduce overfitting both because updating on larger sets moves the model more quickly to convergence, and estimating distributional parameters on larger sets leads to less error.

### 3.2. Linear Scaling

**Algorithm 4** Linear Scaling (LS)

1: Set

$$\text{LS}[f](x) := \text{expit}(\alpha^* + \beta^* \text{logit}\, f(x)), \quad \text{with } (\alpha^*, \beta^*) = \underset{\alpha, \beta}{\arg\min}\, \text{MSE}(\hat{f}). \tag{13}$$

Since Algorithm 3 operates over conditioning events $S_{p,g}$ over which the current model takes constant value, it is reasonable to apply constant patches in the bins that are independent of the model value. However, when we start using conditioning events $S^{\tau}_{p,g}$ like in (12), the model is no longer constant subject to the conditioning events, and it is reasonable to explore alternative updates that depend on the model value. It is immediate to extend the patches in step 4 of Algorithm 3 to linear patches of the form $h_{t+1}(x) = \alpha + \beta f_t(x)$

for $x \in S_{p_t,g_t}$, without affecting the results of Theorem 2.10, so long as we choose $\alpha$ and $\beta$ so as to minimize the squared error of the model. In fact, if $\alpha = \Delta_{p_t,g_t}(f_t)$ and $\beta = 1$, we recover the patch in step 4. This works because the analysis of Algorithm 3 involves showing that the MSE of the model decreases at every step; by minimizing MSE over a model class that can represent the patches used in Step 4 of the algorithm, the analyzed convegence only becomes more rapid. In practice, to avoid clipping $f_t$ between 0 and 1, we make use of the logit and expit (a.k.a. sigmoid) functions, which respectively map $f$ to an unconstrained domain, and then map a linear transformation of it between 0 and 1. Concretely, we replace step 4 in Algorithm 3 with

$$h_{t+1}(x) := \begin{cases} \mathrm{LS}[f_t](x) & \text{if } x \in S^{\tau_t}_{p_t,g_t}(f_t) \\ f_t(x) & \text{otherwise,} \end{cases} \tag{14}$$

where $\mathrm{LS}[f]$ is defined in (13). When applied as a standalone calibration method over the whole input space $\mathcal{X}$, Algorithm 4 is related to similar methods such as Platt scaling (Platt et al., 1999) and temperature scaling (Guo et al., 2017). Instead, we apply it to the conditioning events selected in Algorithm 3 with the goal of multicalibration.

### 3.3. Early stopping

Algorithm 3 is an iterative algorithm that builds a model whose complexity grows with the number of iterations it runs for. Hence a natural heuristic for mitigating overfitting is to implement an early stopping rule. To do so, we initially partition the available data into calibration and validation sets. We then halt the algorithm when the MSE on the validation set ceases to decrease. The rationale of this is supported by Theorem 2.10, which establishes that the MSE must decrease on the calibration data in every round. A lack of MSE reduction signifies potential overfitting, rendering the MSE a meaningful loss function for early stopping.

Another intuitive criterion for early stopping is to assess whether the probability mass of the conditioning event selected by the algorithm for an update exceeds a specified threshold. Recognizing that patches on small conditioning events may potentially compromise the algorithm's generalization ability, we choose to halt the process whenever a conditioning event of insufficient size is chosen for patching.

Algorithm 5 combines the strategies discussed in Section 3.

**Algorithm 5** Iterative Grouped Linear Binning (IGLB)

Let $t = 0$, $f_0 := f'$, $\varepsilon > 0$. Split $\mathcal{D}$ into $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{val}}$.
**while** True **do**
  Set $(p_t, g_t, \tau_t)$ as in (12).
  **Break** if $\mathbb{P}_{\mathcal{D}_{\text{calib}}}\left(S^{\tau_t}_{p_t,g_t}(f_t)\right) < \varepsilon$.
  Set $h_{t+1}(x)$ as in (14).
  **Break** if $\mathrm{MSE}_{\mathcal{D}_{\text{val}}}(h_{t+1}) \geq \mathrm{MSE}_{\mathcal{D}_{\text{val}}}(f_t)$.
  Set $f_{t+1} := h'_{t+1}$.
**end while**

## 4. Application to Hallucination Detection

In this Section we apply the multicalibration techniques developed in Sections 2.2 and 3 to the problem of hallucination detection in LLMs. Our goal is to find a model $f(x)$ which produces (multi)calibrated confidence scores for the probability that a prompt/completion pair $x$ does not correspond to a "hallucination". As discussed, the key problems to solve are what to choose as the "initial" scoring model $f(x)$ and to determine what the groups are for data corresponding to context/completion pairs.

### 4.1. The Initial Scoring Model

Several methods to score hallucinations have been proposed in the literature. While the better the initial scoring model $f(x)$, the better we can expect our final results to be, we remark that our methodology can be applied on top of any scoring method. In this paper, we study 3 different scoring methods proposed in prior work to provide our initial score function $f(x)$, which we refine using multicalibration procedures. Note that all of the scoring functions described below are heuristics and have no calibration guarantees on their own; it is the multicalibration procedure that we use to post-process them that will endow them with guarantees.

**True/False softmax score.** This method employs an LLM to score the correctness of a generated answer for a specific question by asking the model whether the answer is correct and prompting it to respond with either True or False exclusively. Let $s(x)$ represent the softmax computed from the logits of the next token to be generated, and $s_k(x)$ denote the component corresponding to token $k$ (Kadavath et al., 2022). The confidence in the answer True, given the possible answers True and False, is defined as $f(x) := \frac{s_{\text{True}}(x)}{s_{\text{True}}(x)+s_{\text{False}}(x)}$.

**Inverse perplexity score.** In this approach, we use an LLM to compute the output logits for a generated answer to a specific question. The function $f(x)$ is then set as the inverse perplexity of the generated answer, represented as $f(x) := \exp\left(\frac{1}{T-T_0}\sum_{t=T_0+1}^{T} \log p(x_t|x_{:t-1})\right)$, where $x$ is defined as the concatenation of question tokens $x_{1:T_0}$ and answer tokens $x_{T_0+1:T}$ (Jelinek et al., 1977).

**Multiple-choice softmax score.** This approach utilizes an LLM to assess the confidence associated with each potential answer by analyzing the output logits and outputs a score by selecting the maximum confidence amongst the available choices (Kadavath et al., 2022). To elaborate, if $s(x)$ represents the softmax for the upcoming token generation, and $s_{A_k}(x)$ signifies the component corresponding to the $k$-th answer within the set of possible answers $\{A_k\}_{k=1}^K$, the score is defined as $f(x) := \max_{k=1,\ldots,K} \frac{s_{A_k}(x)}{\sum_{k'=1}^{K} s_{A_{k'}}(x)}$.

### 4.2. The Grouping Strategy

Multicalibration is a guarantee parameterized by groups, and so it is important to identify "groups" of prompt/completion pairs $x$ that are correlated with the probability that the completion is a hallucination. When these groups effectively capture features in the data that are associated with increased likelihood of hallucination, their incorporation can lead to substantial improvements in both the algorithm's accuracy and calibration. Multicalibration techniques do not require that the groups are disjoint — the only requirement is that, given a prompt/completion pair $x$, we are able to identify at test time which groups $x$ is a member of. This gives us the freedom to define groups from arbitrary features of the prompt, information about the user, etc. — so long as we have the ability to determine this information in deployment. Here we discuss two strategies for defining groups.

**Clustering.** A natural strategy is to find semantically meaningful *clusters* of prompts within some embedding space. The clustering can use information not only about the prompt, but also about the completion and the scoring model. To the extent that the identified clusters turn out to correlate with the likelihood of hallucination, multicalibrating with respect to groups defined by the identified clusters will improve the underlying scoring function. Off-the-shelf text encoders and soft-clustering methods can be applied in this setting (Devlin et al., 2018; Hastie et al., 2009).

**Annotations.** Another approach to forming groups involves using the LLM to annotate prompt/completion pairs: i.e. re-prompting the LLM with yes-or-no questions whose answers will define the groups. For example, we can ask "Does the following question require at least high school level knowledge?", "Does the following prompt have to do with mathematics?", "Is the following prompt ambiguous?" etc. In general, since the groups need not be disjoint, *any* collection of questions can be used to induce a collection of groups. An LLM can annotate the generated text with True/False assessments, indicating whether it exhibits a set of predefined characteristics. A strength of this approach is that it provides easily human interpretable groups, and is easily extensible compared to clustering strategies. A disadvantage is that it leads to higher computational cost and latency at deployment time, and the quality of the annotations may vary with the LLM.

## 5. Experiments

We conduct a comprehensive experimental comparison of the methodologies introduced in Sections 2, 3, 4.1 and 4.2.

### 5.1. Setup

We conduct experiments on a range of question answering datasets, namely BigBench (Ghazal et al., 2013), MMLU (Hendrycks et al., 2020), OpenBookQA (Mihaylov et al., 2018), TruthfulQA (Lin et al., 2021), MathQA (Amini et al., 2019), and TriviaQA (Joshi et al., 2017). These datasets enable us to assess the methods across a heterogeneous collection of queries over which the probability of hallucination varies substantially. We assess the outcomes using several state-of-the-art LLMs, namely StableBeluga-13B (Touvron et al., 2023; Mukherjee et al., 2023), Flan-T5-base (Chung et al., 2022), Bloomz-7b1 (Muennighoff et al., 2022), and Mistral-7B-v0.1 (Jiang et al., 2023). The goal is to provide a comprehensive understanding of how these methods perform across several datasets and LLMs.

| gASCE | uncalib. | IGLB | IGHB | GCULR | HB | LS |
|---|---|---|---|---|---|---|
| **Business** | 0.0645 | **0.0068** | 0.0189 | 0.0083 | 0.01 | 0.0083 |
| **Computer Sc.** | 0.0824 | 0.0254 | 0.035 | 0.0364 | **0.0241** | 0.0366 |
| **Engineering** | 0.1331 | **0.0523** | 0.0676 | 0.0564 | 0.0679 | 0.0562 |
| **Ethics** | 0.1775 | **0.0189** | 0.0754 | 0.0215 | 0.0703 | 0.0214 |
| **History** | 0.024 | 0.0195 | **0.0178** | 0.025 | 0.0239 | 0.0251 |
| **Law** | 0.1263 | **0.0085** | 0.0422 | 0.0096 | 0.0477 | 0.0096 |
| **Mathematics** | 0.1586 | **0.0231** | 0.0555 | 0.0254 | 0.0264 | 0.0252 |
| **Medicine** | 0.0623 | **0.0064** | 0.0198 | 0.0069 | 0.0547 | 0.007 |
| **Miscellaneous** | 0.0257 | 0.03 | **0.0204** | 0.0321 | 0.0349 | 0.0322 |
| **Philosophy** | 0.0704 | **0.0181** | 0.0312 | 0.0207 | 0.028 | 0.0208 |
| **Political Sc.** | 0.0793 | 0.0439 | 0.0425 | 0.0474 | **0.0223** | 0.0473 |
| **Psychology** | 0.0445 | 0.0118 | 0.0144 | 0.0119 | **0.0104** | 0.0119 |
| **Religion** | 0.0888 | 0.0643 | 0.0808 | 0.0674 | **0.033** | 0.0678 |
| **Science** | 0.0923 | **0.0056** | 0.0244 | 0.0076 | 0.0075 | 0.0076 |
| **Security** | 0.1492 | **0.0237** | 0.0845 | 0.0377 | 0.0329 | 0.0377 |
| **Social Sc.** | 0.0707 | **0.0127** | 0.0296 | 0.0203 | 0.0226 | 0.0204 |

*Table 1.* We report the gASCE for each of the true MMLU topics, on average over different LLMs. An LLM annotation strategy is used in multicalibration methods for grouping. All methods improve the gASCE compared to before calibration. IGLB achieves best results on most groups. In particular, it performs better than GCULR on gASCE, since the first guarantees multicalibration, while the second only group-conditional unbiasedness.

**Labeling the data.** Details in Appendix F.

**Scoring.** Details in Appendix G.

**Grouping.** Details in Appendix H.

### 5.2. Results

We conduct a comparative analysis by comparing the initial scoring functions (without any post-processing for calibration) against the same scoring functions post-processed for calibration using several algorithms: Algorithm 1 (HB), Algorithm 4 (LS), the logistic regression version of Algorithm 2 (GCULR), Algorithm 3 (IGHB), and Algorithm 5 (IGLB). HB and LS aim for standard (marginal) calibration and do not use any grouping strategy. GCULR produces a model satisfying group-conditional unbiasedness but not necessarily multicalibration. IGHB and IGLB produce multicalibrated models. Compared to IGHB, IGLB implements all of the modifications discussed in Section 3 to mitigate overfitting. All experiments in this section employ the True/False softmax score described in Section 4.1. For results using other scoring methods, please refer to Appendix I.

| MSE | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **uncalib.** | 0.3242 (0.0201) | 0.3045 (0.0315) | 0.2608 (0.0037) | 0.4762 (0.135) | 0.3767 (0.0817) | 0.2802 (0.029) |
| **IGLB** | **0.2416** (0.0027) | 0.2254 (0.0084) | 0.236 (0.0091) | **0.2016** (0.0437) | **0.1727** (0.0047) | **0.1974** (0.0308) |
| **IGHB** | 0.2588 (0.0157) | 0.2517 (0.0138) | 0.2517 (0.0062) | 0.3051 (0.0147) | 0.1898 (0.0083) | 0.2078 (0.0299) |
| **GCULR** | 0.2432 (0.0028) | **0.2239** (0.0083) | **0.2354** (0.009) | 0.2047 (0.0471) | 0.1728 (0.0047) | 0.1976 (0.0306) |
| **HB** | 0.2444 (0.0018) | 0.23 (0.009) | 0.2357 (0.0094) | 0.2043 (0.041) | 0.1728 (0.0047) | 0.2026 (0.0334) |
| **LS** | 0.2459 (0.0024) | 0.2281 (0.0093) | 0.236 (0.0091) | 0.2036 (0.0457) | **0.1727** (0.0047) | 0.2008 (0.031) |
| **ACC.** | **BigBench** | **MMLU** | **OpenBookQA** | **TruthfulQA** | **MathQA** | **TriviaQA** |
| **uncalib.** | 0.4815 (0.0443) | 0.4961 (0.0258) | 0.5506 (0.037) | 0.3333 (0.1219) | 0.3131 (0.0975) | 0.5766 (0.0372) |
| **IGLB** | **0.5691** (0.0114) | 0.634 (0.0325) | 0.5933 (0.0356) | 0.6871 (0.1158) | **0.7779** (0.0085) | 0.7023 (0.083) |
| **IGHB** | 0.5462 (0.0142) | 0.5858 (0.0047) | 0.5711 (0.0476) | 0.4843 (0.0655) | 0.7421 (0.0162) | 0.6781 (0.0899) |
| **GCULR** | 0.5548 (0.0128) | **0.6381** (0.0313) | **0.5979** (0.0282) | 0.6777 (0.128) | **0.7779** (0.0085) | **0.7037** (0.0812) |
| **HB** | 0.5613 (0.0091) | 0.6274 (0.0353) | 0.5933 (0.0447) | **0.6997** (0.1034) | **0.7779** (0.0085) | 0.6944 (0.0904) |
| **LS** | 0.5582 (0.0169) | 0.6299 (0.034) | 0.5925 (0.0347) | 0.6698 (0.1345) | **0.7779** (0.0085) | 0.6953 (0.0896) |

*Table 2.* MSE and accuracy metrics are presented for all methods across various datasets, with results displayed as the mean and standard deviation (in brackets) derived from the values produced by four LLMs. Our findings highlight the superiority of multicalibration methods, specifically IGLB and GCULR, over alternative approaches across all datasets. In particular, IGLB demonstrates a significant performance advantage over IGHB, emphasizing the effectiveness of the overfitting remedies proposed in Section 3.

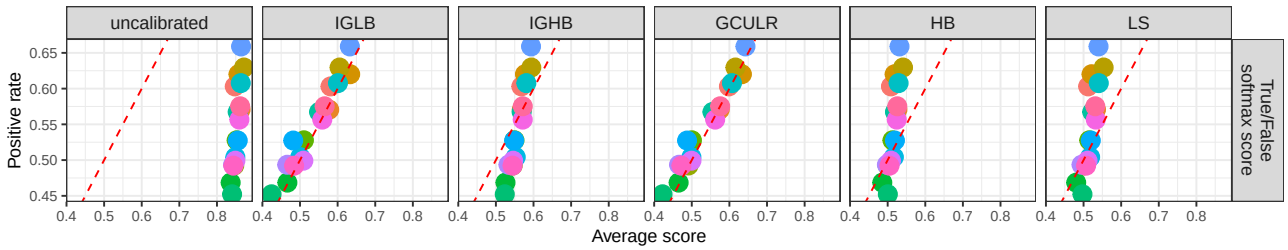


*Figure 2.* Average scores against accuracies across various clusters, for each method, on MMLU for StableBeluga-13B. Colors represent the groups, and the size of the points reflects their size. Multicalibration methods exhibit significantly superior alignment with the diagonal compared to standard calibration methods. In agreement with the results in Table 2, IGLB and GCULR stand out as the top performers.

In Table 2, we present comprehensive statistics, including the mean and standard deviation (in brackets) of MSE and binary classification accuracy across the different LLMs. These metrics are analyzed across all methods and datasets. When evaluating binary classification accuracy, binary predictions are derived from the scoring function by applying a threshold of $\frac{1}{2}$ to the scores. This is the threshold that maximizes classification accuracy whenever the scoring function is calibrated. The results presented here use the clustering grouping strategy. For a discussion of the results with the annotation grouping strategy, please refer to Appendix L.

Results show that IGLB and GCULR consistently outperform across all datasets, respectively leading on MSE and accuracy. All of the models that are post-processed for calibration out-perform the initial scoring function, sometimes substantially. This underscores the importance of calibration post-processing in enhancing detection capabilities.

The results also confirm that IGHB, in its original form, is prone to overfitting. However, the modifications detailed in Section 3 which are incorporated into IGLB, significantly enhance performance. Further insights into the specific effects of these changes are explored in Appendix J.

Standard calibration methods such as HB and LS perform well compared to the initial scoring function, yet consistently underperform IGLB and GCULR. It's noteworthy that HB achieves the best results only on OpenBookQA, possibly because of the small dataset size which causes the more complex models to overfit.

Figure 2 illustrates average confidence scores plotted against the fraction of positive labels across various groups (each corresponding to a different color in the plot), to evaluate group-wise calibration. Once again we see IGLB and GCULR standing out as the top performers (represented as alignment with the diagonal). See also Appendix K.

Table 6 provides further evidence that IGLB outperforms other methods on the gASCE evaluated using true MMLU topics. More details about this experiment in Appendix L.

## 6. Conclusions

In this paper we introduce multicalibration to confidence scoring for LLMs, and develop several new techniques for both generating groups of prompts to multicalibrate with respect to, as well as new multicalibration algorithms which

have improved practical performance. We show that when applied to existing scoring functions from the literature, our methods substantially improve both the error and calibration of the scores. What we have presented is an extensible framework, and so there is a clear pathway to improvement via new grouping strategies that are both semantically meaningful and correlated with LLM performance.

## Impact Statement

By calibrating the confidence associated with text generated by LLMs, this paper contributes to enhance trustworthiness in applications ranging from customer service interactions, to content creation and educational platforms. Calibration not only ensures more reliable and contextually appropriate responses but also mitigates the risks associated with biased or inappropriate content, thereby aligning with ethical considerations in AI development. The impact extends to critical sectors such as healthcare and finance, where the reliability of AI-generated information is of paramount importance. Furthermore, the calibration process facilitates model explainability, offering insights into decision-making mechanisms and promoting transparency in AI systems. In essence, the calibration of hallucination in LLMs is a pivotal step toward fostering responsible, trustworthy, and ethically sound AI technologies. It is important to note, however, that calibration methods should be used only with an understanding of their limitations. In our case, the provable calibration guarantees are designed to hold on prompts that are distributed as those in our calibration set are, and do not hold for adversarially generated prompts.

## References


Amini, A., Gabriel, S., Lin, P., Koncel-Kedziorski, R., Choi, Y., and Hajishirzi, H. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019.

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H.-S., Fernández, R., Plank, B., Sennrich, R., Zerva, C., and Aziz, W. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*, 2023.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Chen, J. and Mueller, J. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*, 2023.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., and Xiao, Y. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 245–255, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Deutschmann, N., Alberts, M., and Martínez, M. R. Conformal autoregressive generation: Beam search with coverage guarantees. *arXiv preprint arXiv:2309.03797*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Duan, J., Cheng, H., Wang, S., Wang, C., Zavalny, A., Xu, R., Kailkhura, B., and Xu, K. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*, 2023.

Elaraby, M., Lu, M., Dunn, J., Zhang, X., Wang, Y., and Liu, S. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*, 2023.

Eyuboglu, S., Varma, M., Saab, K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Ré, C. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Friel, R. and Sanyal, A. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*, 2023.

Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., and Jacobsen, H.-A. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pp. 1197–1208, 2013.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Globus-Harris, I., Harrison, D., Kearns, M., Roth, A., and Sorrell, J. Multicalibration as boosting for regression. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11459–11492. PMLR, 2023. URL https://proceedings.mlr.press/v202/globus-harris23a.html.

Gopalan, P., Hu, L., Kim, M. P., Reingold, O., and Wieder, U. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022a.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. Omnipredictors. *Leibniz international proceedings in informatics*, 215, 2022b.

Gopalan, P., Kim, M. P., Singhal, M. A., and Zhao, S. Low-degree multicalibration. In *Conference on Learning Theory*, pp. 3193–3234. PMLR, 2022c.

Gopalan, P., Hu, L., Kim, M. P., Reingold, O., and Wieder, U. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023a.

Huang, Y., Song, J., Wang, Z., Chen, H., and Ma, L. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023b.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Johnson, N., Cabrera, Á. A., Plumb, G., and Talwalkar, A. Where does my model underperform? a human evaluation of slice discovery algorithms. *arXiv preprint arXiv:2306.08167*, 2023.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kalai, A. T. and Vempala, S. S. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Kohavi, R., Wolpert, D. H., et al. Bias plus variance decomposition for zero-one loss functions. In *ICML*, volume 96, pp. 275–283, 1996.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

Li, X. and Li, J. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., and Dolan, B. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*, 2021.

Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., and Zhang, Q. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*, 2023.

Luo, J., Xiao, C., and Ma, F. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*, 2023.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4, 2023.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Noarov, G., Ramalingam, R., Roth, A., and Xie, S. High-dimensional prediction for sequential decision making. *arXiv preprint arXiv:2310.17651*, 2023.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., and Cohen, J. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Roth, A. Uncertain: Modern topics in uncertainty estimation, 2022.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Varshney, N., Yao, W., Zhang, H., Chen, J., and Yu, D. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

Verma, S., Tran, K., Ali, Y., and Min, G. Reducing llm hallucinations using epistemic neural networks. *arXiv preprint arXiv:2312.15576*, 2023.

Xiao, Y. and Wang, W. Y. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.

Yang, Q., Ravikumar, S., Schmitt-Ulms, F., Lolla, S., Demir, E., Elistratov, I., Lavaee, A., Lolla, S., Ahmadi, E., Rus, D., et al. Uncertainty-aware language modeling for selective question answering. *arXiv preprint arXiv:2311.15451*, 2023.

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., and Yuan, L. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Ye, H., Liu, T., Zhang, A., Hua, W., and Jia, W. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 609–616, 2001.

Zecchin, M., Park, S., and Simeone, O. Forking uncertainties: Reliable prediction and model predictive control with sequence models via conformal risk control. *arXiv preprint arXiv:2310.10299*, 2023.

Zhang, J., Li, Z., Das, K., Malin, B. A., and Kumar, S. SAC$^3$: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *arXiv preprint arXiv:2311.01740*, 2023a.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Zhao, T., Wei, M., Preston, J. S., and Poon, H. Automatic calibration and error correction for large language models via pareto optimal self-supervision. *arXiv preprint arXiv:2306.16564*, 2023.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., and Wen, J.-R. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

# Appendix

## A. On the expected calibration error

In our notation, the expected calibration error (ECE) is defined as

$$\text{ECE}(f) := \sum_{i=1}^{m} \mathbb{P}(f(X) \in B_i)\, \mathbb{E}_{\mathcal{D}}\Big[|\text{Accuracy}(f) - \text{Confidence}(f)| \Big| f(X) \in B_i\Big],$$

where Accuracy$(f)$ is defined as $\mathbb{P}(Y = \mathbb{1}[f(X) \geq \frac{1}{2}])$, $B_i := [\frac{i-1}{m}, \frac{i}{m}]$, and Confidence denotes the probability of the predicted label, that is

$$\text{Confidence}(f) := \max\{1 - f(x), f(x)\}.$$

## B. Proof of Proposition 2.4

We have

$$\begin{aligned}
\text{MSE}(f) &= \mathbb{E}_{\mathcal{D}}[(Y - f(X))^2] \\
&= \mathbb{E}_P[\mathbb{E}_{\mathcal{D}}[(Y - P)^2 | f(X) = P]] \\
&= \mathbb{E}_P[\mathbb{E}_{\mathcal{D}}[(Y - \mathbb{E}_{\mathcal{D}}[Y|f(X) = P] + \mathbb{E}_{\mathcal{D}}[Y - P|f(X) = P])^2 | f(X) = P]] \\
&= \mathbb{E}_P[\text{Var}_{\mathcal{D}}(Y|f(X) = P)] + \mathbb{E}_P[\mathbb{E}_{\mathcal{D}}[Y - P|f(X) = P]^2] \\
&= \mathbb{E}_P[\text{Var}_{\mathcal{D}}(Y|f(X) = P)] + \text{ASCE}(f).
\end{aligned}$$

## C. Proof of Theorem 2.5

First, we show that ASCE$(\hat{f}) = 0$. Notice that, by construction of $\hat{f}$, any element $p$ in the image of $\hat{f}$ must either belong to $[\frac{1}{m}]$ or it must be such that $p = \hat{f}(x) = f'(x) + \Delta_{p'}(f')$, for some $p' \in [\frac{1}{m}]$. In both cases, it is immediate to check that $\Delta_p(\hat{f}) = 0$. It follows that

$$\text{ASCE}(\hat{f}) = \mathbb{E}_{P \in \text{Im}(\hat{f})}[\Delta_P^2(\hat{f})] = 0.$$

We now prove the decrease in MSE. We have

$$\begin{aligned}
\text{MSE}(\hat{f}) &= \mathbb{E}_{\mathcal{D}}[(Y - \hat{f}(X))^2] \\
&= \mathbb{E}_{\mathcal{D}}[(Y - f'(X) + f'(X) - \hat{f}(X))^2] \\
&= \text{MSE}(f') - \sum_{p \in [\frac{1}{m}]} \mathbb{P}_{\mathcal{D}}(B_p) \Delta_p^2(f').
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\text{MSE}(f') &= \mathbb{E}_{\mathcal{D}}[(Y - f'(X))^2] \\
&= \mathbb{E}_{\mathcal{D}}[(Y - f(X) + f(X) - f'(X))^2] \\
&= \text{MSE}(f) + \sum_{p \in [\frac{1}{m}]} \mathbb{P}_{\mathcal{D}}(B_p) \mathbb{E}_{\mathcal{D}}[(f(X) - f'(X))^2 + 2(Y - f(X))(f(X) - f'(X)) | B_p].
\end{aligned}$$

Notice that $\mathbb{E}_{\mathcal{D}}[(f(X) - f'(X))^2 | B_p] < \frac{1}{4m^2} = \frac{\alpha^2}{4}$. Furthermore,

$$\mathbb{E}_{\mathcal{D}}\left[(Y - f(X))(f(X) - f'(X)) | B_p\right] \leq \mathbb{E}_{\mathcal{D}}\left[|Y - f(X)| \big| B_p\right] \mathbb{E}_{\mathcal{D}}\left[|f(X) - f'(X)| \big| B_p\right] < 1 \cdot \frac{1}{m} = \alpha.$$

It follows that MSE$(f') <$ MSE$(f) + \frac{\alpha^2}{4} + \alpha$, which concludes the proof.

## D. Proof of Theorem 2.7

For each $g \in \mathcal{G}$, we have

$$\begin{aligned}\frac{\partial}{\partial \lambda_g}\text{MSE}(\hat{f}) &= 2\mathbb{E}_{\mathcal{D}}[\frac{\partial}{\partial \lambda_g}\hat{f}(X)(\hat{f}(X) - Y)] \\ &= 2\mathbb{E}_{\mathcal{D}}[g(X)(\hat{f}(X) - Y)] \\ &= 2\mathbb{E}_{\mathcal{D}}[\hat{f}(X) - Y | g(X) = 1],\end{aligned}$$

where the last line follows from $g(X) \in \{0, 1\}$. Then $\frac{\partial}{\partial \lambda_g}\text{MSE}(\hat{f}) = 0$ if and only if $\mathbb{E}_{\mathcal{D}}[Y - \hat{f}(X)|g(X) = 1] = 0$, hence the model $\hat{f}$ parametrized by the $\{\lambda_g\}_{g\in\mathcal{G}}$ that minimizes the MSE must also satisfy group-conditional unbiasedness.

The result also holds for the cross-entropy loss

$$\text{CrossEntropy}(\hat{f}) := \mathbb{E}_{\mathcal{D}}[Y \log \hat{f}(X) + (1 - Y)\log(1 - \hat{f}(X))].$$

Indeed,

$$\begin{aligned}\left|\frac{\partial}{\partial \lambda_g}\text{CrossEntropy}(\hat{f})\right| &= \left|\mathbb{E}_{\mathcal{D}}\left[\frac{\partial}{\partial \lambda_g}\hat{f}(X)\left(\frac{Y}{\hat{f}(X)} - \frac{1-Y}{1-\hat{f}(X)}\right)\right]\right| \\ &= \left|\mathbb{E}_{\mathcal{D}}\left[\frac{Y - \hat{f}(X)}{\hat{f}(X)(1-\hat{f}(X))}\Big| g(X) = 1\right]\right| \\ &\geq 4\left|\mathbb{E}_{\mathcal{D}}[Y - \hat{f}(X)|g(X) = 1]\right|.\end{aligned}$$

Hence a model $\hat{f}$ parametrized by the $\{\lambda_g\}_{g\in\mathcal{G}}$ that minimizes the cross-entropy loss must also satisfy group-conditional unbiasedness.

## E. Proof of Proposition 3.1

We want to show that the following two conditions are equivalent:

$$\mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = p] = 0 \quad \forall\, p \in [0, 1] \qquad (1)$$

$$\mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) \leq p] = 0 \quad \forall\, p \in [0, 1] \qquad (2)$$

First, notice that

$$\mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) \leq p] = \frac{\int_0^p \mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = v]\, p_{f(X)}(v)\, dv}{\mathbb{P}(f(X) \leq p)}.$$

It follows that if (1) holds, then (2) must hold, since the expression within the integral above is 0. Vice versa, if (2) holds, then we must have

$$\int_{p_1}^{p_2} \mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = v]\, p_{f(X)}(v)\, dv = 0 \quad \forall\, p_1, p_2 \in [0, 1].$$

Because $\mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = v]$ is continuous in $v$ by assumption, it follows that if there existed $p^* \in [0, 1]$ such that $\mathbb{E}_{\mathcal{D}}[Y - f(X)|f(X) = p^*] \neq 0$, then there would exist a small enough interval $[p_1^*, p_2^*] \ni p^*$ where the function does not change its sign. Hence the integral above over this interval would not be zero, which would contradict (2).

## F. Labeling the data

To calibrate a scoring function, we need a calibration dataset which consists of a collection of questions, answers, and binary labels for each answer corresponding to whether or not the answer is "correct" or not. While BigBench, MMLU, OpenBookQA, TruthfulQA, and MathQA conveniently contain questions, multiple answers per question, and binary labels

for each answer indicating correctness, TriviaQA poses a challenge by providing only a set of correct answers for each question. To overcome this limitation, we construct labelled data for TriviaQA as follows. For each TriviaQA question, we prompt an LLM to generate four different answers. Subsequently, binary labels are assigned to these generated answers based on word overlap with the answer key provided in the dataset. After this processing, TriviaQA becomes a multiple choice question answering problem, where the multiple choice options depend on the LLM under study. The data is then randomly split into calibration and testing sets, with an 80/20 split.

## G. Scoring

For each LLM and dataset, we experiment with each of the three initial scoring models outlined in Section 4.1. Given that our datasets consist of multiple answers per question, we select the answer with the highest score and use this score as the output of the initial model $f$.

## H. Grouping setup

To form groups, we follow the methods described in Section 4.2. For the clustering approach, text embeddings were obtained using UAE-Large-V1 (Li & Li, 2023), reduction to 20-dimensional embeddings was performed using UMAP (McInnes et al., 2018), and a Gaussian mixture model was fitted on both the embeddings and one of the types of LLM scores for the most likely answer. The Bayesian Information Criterion was used to select the number of groups. We also evaluated alternative dimensionality reduction techniques such as t-SNE (Van der Maaten & Hinton, 2008) and PCA, as well as the use of other text encoders, obtaining conclusions similar to those described in the paper. Our methodology aligns with clustering approaches in the blindspot discovery literature (Eyuboglu et al., 2022; Johnson et al., 2023), where clustering ensures that items in each group are semantically related and exhibit comparable classification accuracy.

For the method that relies on annotations, our goal was to design a comprehensive taxonomy that would cover the wide range of questions appearing in the datasets. Thus, we gave StableBeluga2 questions from the different datasets and instructed it to create a classification system with groups that intersect, and then further group related categories into areas. The resulting groups for all datasets are reported in Tables 3. For MMLU, where the topic of each question is already provided, we asked the LLM to create a simplified classification system with fewer and non-disjoint categories based on the existing taxonomy. We also instructed it to align the original categories with the new ones and manually refined this mapping. The categories created for MMLU include those reported in Section L. To obtain the annotations, we asked the model "You are an AI designed to categorize questions accurately. The possible categories are as follows: <all possible categories>. Which of these categories does the question <question from dataset> fall into?". We then picked the most likely categories according to the confidence scores produced by the LLM.

## I. Results with other scoring methods

In this section, we present results similar to those in Table 2, employing the inverse perplexity and multi-choice scoring methods introduced in Section 4.1. Across all datasets, the findings consistently validate the earlier observations, indicating that IGLB and GCULR perform generally better than other methods.

## J. Ablation study

In this section, we examine in isolation the impacts resulting from the modifications proposed in Section 3. Specifically, we conduct a comparative analysis between IGHB and IGLB against two distinct variants:

- **IGHB**$^{\tau}$: This variant mirrors IGHB but leverages lower and upper sets $S^{\tau}_{p,g}$ for bins, as expounded upon in Section 3.1.
- **IGHB**$^{\text{LS}}$: This variant employs the linear scaling patching strategy discussed in Section 3.2 instead of the constant shift patch in Algorithm 3.

The results in Table 5 consistently reveal that IGLB, encompassing all the proposed changes to address overfitting in Section 3, consistently outperforms its variants, which either lack or only partially incorporate the proposed alterations. The lower and upper-level binning scheme, as introduced in Section 3.1, emerges as the primary driver of improvement, with IGHB$^{\tau}$ achieving nearly comparable results to IGLB. Conversely, the implementation of linear scaling in IGHB$^{\text{LS}}$, without the

| Category | Area |
|---|---|
| Humanities | Subject-Based Knowledge |
| Social Sciences | Subject-Based Knowledge |
| Natural Sciences | Subject-Based Knowledge |
| Formal Sciences | Subject-Based Knowledge |
| Professional Knowledge | Subject-Based Knowledge |
| Basic Arithmetic | Mathematical Reasoning |
| Algebra | Mathematical Reasoning |
| Geometry | Mathematical Reasoning |
| Advanced Mathematics | Mathematical Reasoning |
| Statistical and Probabilistic Reasoning | Mathematical Reasoning |
| Problem Solving | Logical and Critical Thinking |
| Inferential Reasoning | Logical and Critical Thinking |
| Analytical Reasoning | Logical and Critical Thinking |
| Common Misconceptions | Factual Accuracy and Misconceptions |
| Fact-Checking | Factual Accuracy and Misconceptions |
| Controversial and Sensitive Topics | Factual Accuracy and Misconceptions |
| Cultural Literacy | General Knowledge and Trivia |
| Historical Facts | General Knowledge and Trivia |
| Scientific Facts | General Knowledge and Trivia |
| Real-World Application | Applied Knowledge |
| Hypothetical Scenarios | Applied Knowledge |
| Cross-Disciplinary Questions | Interdisciplinary |
| Integrative Reasoning | Interdisciplinary |

*Table 3.* LLM-generated taxonomy of questions appearing in the datasets described in Section 5.1.

alteration to the bins, results in poorer performance than IGHB alone. This observation is unsurprising, as within standard level set bins, the model remains almost constant, and linear scaling patches merely amplify the risk of overfitting.

## K. Scatter plots

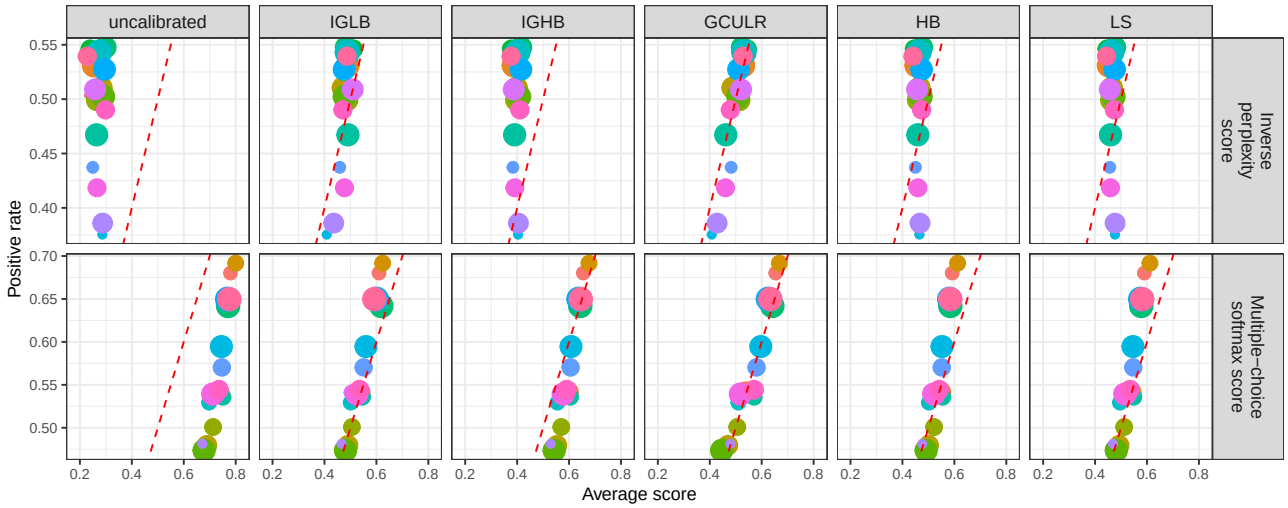


*Figure 3.* The average scores against the accuracy across various clusters, for each calibration method, and for inverse perplexity and multiple-choice softmax scores on MMLU and StableBeluga-13B. Conclusions are similar to those derived for Figure 2.

Figure 3 provides insights analogous to Figure 2, but using inverse perplexity and multiple-choice softmax scores. All calibration methods provide post-processed scores that align significantly better with the diagonal than the intial scores

**Inverse perplexity score**

| MSE | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **uncalib.** | 0.3506 (0.0104) | 0.2746 (0.0133) | 0.3506 (0.1042) | 0.22 (0.0288) | 0.2268 (0.0179) | 0.2294 (0.058) |
| **IGLB** | 0.2441 (0.0033) | 0.2183 (0.0155) | 0.2229 (0.0417) | 0.1857 (0.0232) | **0.1991** (0.0223) | 0.1863 (0.0352) |
| **IGHB** | 0.2593 (0.0083) | 0.2287 (0.0219) | 0.2339 (0.0447) | 0.2039 (0.0331) | 0.2062 (0.0218) | 0.1847 (0.0307) |
| **GCULR** | **0.2439** (0.0045) | **0.2181** (0.0152) | **0.2224** (0.0403) | **0.1844** (0.0249) | 0.2001 (0.0234) | **0.1795** (0.0307) |
| **HB** | 0.2456 (0.0024) | 0.2212 (0.0174) | 0.2213 (0.0409) | 0.1912 (0.0262) | 0.1995 (0.0226) | 0.1876 (0.0338) |
| **LS** | 0.2473 (0.0016) | 0.223 (0.0161) | 0.2226 (0.0418) | 0.1898 (0.0207) | 0.2004 (0.0236) | 0.1917 (0.0351) |

| ACC. | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **uncalib.** | 0.5278 (0.0315) | 0.6495 (0.0573) | 0.6156 (0.1184) | 0.6997 (0.0888) | 0.7087 (0.0448) | 0.7016 (0.0589) |
| **IGLB** | **0.5516** (0.0152) | **0.659** (0.0499) | **0.6232** (0.1162) | **0.7453** (0.0506) | **0.719** (0.0498) | 0.7178 (0.0714) |
| **IGHB** | 0.5334 (0.0234) | 0.6495 (0.0573) | 0.6156 (0.1184) | 0.706 (0.081) | 0.7087 (0.0448) | 0.7312 (0.0606) |
| **GCULR** | 0.5487 (0.0173) | 0.6573 (0.0509) | 0.6225 (0.115) | **0.7453** (0.0506) | 0.7148 (0.0543) | **0.7502** (0.0707) |
| **HB** | 0.5511 (0.0129) | 0.6497 (0.0586) | 0.6221 (0.1142) | 0.7437 (0.0494) | 0.7177 (0.0511) | 0.7132 (0.0729) |
| **LS** | 0.5493 (0.0151) | 0.6489 (0.0576) | 0.6214 (0.1175) | **0.7453** (0.0506) | 0.7138 (0.0553) | 0.7111 (0.074) |

**Multiple-choice softmax score**

| MSE | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **uncalib.** | 0.2885 (0.0144) | 0.2476 (0.0333) | 0.2203 (0.0418) | 0.2713 (0.0343) | 0.1998 (0.0189) | 0.2359 (0.0279) |
| **IGLB** | **0.238** (0.0016) | 0.2068 (0.0089) | **0.1985** (0.0308) | 0.2058 (0.0259) | **0.1728** (0.0042) | 0.1778 (0.0416) |
| **IGHB** | 0.2513 (0.0028) | 0.2249 (0.0226) | 0.2175 (0.0417) | 0.2661 (0.0361) | 0.1827 (0.0124) | 0.2088 (0.0373) |
| **GCULR** | **0.238** (0.0016) | **0.2053** (0.0107) | 0.1986 (0.0306) | **0.2043** (0.0245) | 0.1729 (0.0041) | **0.1777** (0.0416) |
| **HB** | 0.2407 (0.0019) | 0.2083 (0.01) | 0.2 (0.0305) | 0.2044 (0.0255) | 0.173 (0.0045) | 0.1809 (0.0415) |
| **LS** | 0.242 (0.0018) | 0.2076 (0.0096) | 0.1984 (0.0308) | 0.2041 (0.0244) | 0.1728 (0.0044) | 0.1807 (0.0413) |

| ACC. | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **uncalib.** | 0.5176 (0.034) | 0.5929 (0.0587) | 0.6435 (0.0801) | 0.555 (0.0684) | 0.7339 (0.0516) | 0.6536 (0.0602) |
| **IGLB** | 0.5769 (0.0102) | 0.6821 (0.0148) | 0.6837 (0.0597) | 0.6824 (0.0702) | **0.7762** (0.009) | 0.7389 (0.0891) |
| **IGHB** | 0.5594 (0.0044) | 0.655 (0.0219) | 0.6525 (0.0735) | 0.5566 (0.0693) | 0.7501 (0.028) | 0.6796 (0.0598) |
| **GCULR** | **0.5774** (0.0057) | **0.6859** (0.0155) | **0.688** (0.0564) | 0.6824 (0.0719) | **0.7762** (0.009) | **0.7396** (0.0884) |
| **HB** | 0.5738 (0.0143) | 0.6791 (0.0177) | 0.685 (0.0578) | **0.695** (0.0675) | **0.7762** (0.009) | 0.7363 (0.0917) |
| **LS** | 0.5736 (0.0138) | 0.6797 (0.0165) | 0.6843 (0.0585) | 0.6855 (0.0669) | 0.776 (0.0093) | 0.7355 (0.0924) |

*Table 4.* Comparable outcomes to those presented in Table 2 are reported, utilizing the inverse perplexity and multiple-choice scoring methods detailed in Section 4.1. The overall findings reinforce that IGLB and GCULR consistently outperform other methods across all datasets.

| MSE | BigBench | MMLU | OpenBookQA | TruthfulQA | MathQA | TriviaQA |
|---|---|---|---|---|---|---|
| **IGLB** | **0.2416** (0.0027) | **0.2254** (0.0084) | **0.236** (0.0091) | **0.2016** (0.0437) | **0.1727** (0.0047) | **0.1974** (0.0308) |
| **IGHB**$^{\tau}$ | 0.2428 (0.0024) | 0.2269 (0.0096) | 0.2372 (0.0083) | 0.2043 (0.0451) | 0.173 (0.0049) | 0.1977 (0.0305) |
| **IGHB**$^{\mathrm{LS}}$ | 0.2521 (0.0101) | 0.2597 (0.0206) | 0.2583 (0.0051) | 0.3421 (0.0424) | 0.2 (0.0219) | 0.2243 (0.017) |
| **IGHB** | 0.2588 (0.0157) | 0.2517 (0.0138) | 0.2517 (0.0062) | 0.3051 (0.0147) | 0.1898 (0.0083) | 0.2078 (0.0299) |
| **ACC.** | **BigBench** | **MMLU** | **OpenBookQA** | **TruthfulQA** | **MathQA** | **TriviaQA** |
| **IGLB** | **0.5691** (0.0114) | 0.634 (0.0325) | **0.5933** (0.0356) | **0.6871** (0.1158) | **0.7779** (0.0085) | 0.7023 (0.083) |
| **IGHB**$^{\tau}$ | 0.5592 (0.013) | **0.6346** (0.0347) | 0.5841 (0.0296) | 0.6698 (0.1335) | **0.7779** (0.0085) | **0.7026** (0.0827) |
| **IGHB**$^{\mathrm{LS}}$ | 0.5524 (0.0093) | 0.5497 (0.025) | 0.5519 (0.0404) | 0.4733 (0.0502) | 0.7398 (0.0272) | 0.6498 (0.0614) |
| **IGHB** | 0.5462 (0.0142) | 0.5858 (0.0047) | 0.5711 (0.0476) | 0.4843 (0.0655) | 0.7421 (0.0162) | 0.6781 (0.0899) |

*Table 5.* The results consistently demonstrate that IGLB outperforms its variants in the majority of cases. The principal contributor to these improvements is the lower and upper-level binning scheme implemented in IGHB$^{\tau}$.

before post-processing. Also in this case, multicalibration methods tend to perform better than methods such as HB and LS, which only satisfy calibration.

| gASCE | uncalib. | IGLB | IGHB | GCULR | HB | LS |
|---|---|---|---|---|---|---|
| **Business** | 0.0645 (0.0291) | **0.0068** (0.0069) | 0.0189 (0.0133) | 0.0083 (0.0056) | 0.01 (0.0048) | 0.0083 (0.0056) |
| **Computer Sc.** | 0.0824 (0.0396) | 0.0254 (0.0139) | 0.035 (0.0113) | 0.0364 (0.0066) | **0.0241** (0.0127) | 0.0366 (0.0066) |
| **Engineering** | 0.1331 (0.0213) | **0.0523** (0.0394) | 0.0676 (0.0323) | 0.0564 (0.0072) | 0.0679 (0.0434) | 0.0562 (0.0072) |
| **Ethics** | 0.1775 (0.0865) | **0.0189** (0.0104) | 0.0754 (0.0716) | 0.0215 (0.009) | 0.0703 (0.0899) | 0.0214 (0.0088) |
| **History** | 0.024 (0.0128) | 0.0195 (0.0087) | **0.0178** (0.0047) | 0.025 (0.0074) | 0.0239 (0.0121) | 0.0251 (0.0073) |
| **Law** | 0.1263 (0.0381) | **0.0085** (0.0042) | 0.0422 (0.0254) | 0.0096 (0.003) | 0.0477 (0.0713) | 0.0096 (0.0032) |
| **Mathematics** | 0.1586 (0.0852) | **0.0231** (0.0137) | 0.0555 (0.0119) | 0.0254 (0.0122) | 0.0264 (0.0132) | 0.0252 (0.0121) |
| **Medicine** | 0.0623 (0.0266) | **0.0064** (0.0039) | 0.0198 (0.0122) | 0.0069 (0.0028) | 0.0547 (0.0654) | 0.007 (0.0029) |
| **Miscellaneous** | 0.0257 (0.0091) | 0.03 (0.0269) | **0.0204** (0.0067) | 0.0321 (0.0225) | 0.0349 (0.0257) | 0.0322 (0.0225) |
| **Philosophy** | 0.0704 (0.0285) | **0.0181** (0.0117) | 0.0312 (0.0066) | 0.0207 (0.0074) | 0.028 (0.0076) | 0.0208 (0.0071) |
| **Political Sc.** | 0.0793 (0.0268) | 0.0439 (0.0288) | 0.0425 (0.0082) | 0.0474 (0.0229) | **0.0223** (0.0152) | 0.0473 (0.0228) |
| **Psychology** | 0.0445 (0.0229) | 0.0118 (0.0071) | 0.0144 (0.0032) | 0.0119 (0.0051) | **0.0104** (0.0053) | 0.0119 (0.0051) |
| **Religion** | 0.0888 (0.04) | 0.0643 (0.0337) | 0.0808 (0.0225) | 0.0674 (0.0314) | **0.033** (0.0195) | 0.0678 (0.0316) |
| **Science** | 0.0923 (0.049) | **0.0056** (0.003) | 0.0244 (0.0098) | 0.0076 (0.0015) | 0.0075 (0.0026) | 0.0076 (0.0014) |
| **Security** | 0.1492 (0.0526) | **0.0237** (0.0183) | 0.0845 (0.0388) | 0.0377 (0.0164) | 0.0329 (0.0329) | 0.0377 (0.0163) |
| **Social Sc.** | 0.0707 (0.0326) | **0.0127** (0.0083) | 0.0296 (0.0214) | 0.0203 (0.0109) | 0.0226 (0.0087) | 0.0204 (0.0108) |

*Table 6.* We report the gASCE obtained by each method for each of the true MMLU topics, with True/False softmax scores. An LLM annotation strategy is used in multicalibration methods for grouping. All methods improve the gASCE compared to before calibration. In particular, IGLB achieves best results almost most groups. It is meaningful to notice that, as expected from the theory, IGLB achieves better results than GCULR on gASCE, since the first guarantees multicalibration, while the second only group-conditional unbiasedness.

## L. Results with annotations

The MMLU dataset is organized by different "topics" (e.g. "Engineering", "Science", etc — see Table 6). The quality of a model's responses can differ substantially by topic. Here we use the LLM to attempt to annotate each prompt by topic, and then use these self-annotations as groups in our multicalibration methods. Note that the self-annotations may differ from the "true" groupings in the MMLU dataset because of errors in the LLM annotations or ambiguities. We can nevertheless evaluate the calibration error of each of the methods we experiment with on the *true* topic groupings within the MMLU dataset. In Table 6, we present the mean (for standard deviation, see Appendix L) of the gASCE for each true topic of MMLU. Similar to Table 2, these values are computed across the mentioned LLMs.

We see that all methods offer improvements in calibration error compared to the uncalibrated raw scores. However, we see that the multicalibration methods which make explicit use of the self-annotated group labels substantially out-perform methods that aim for only marginal calibration. Notably, IGLB consistently achieves the lowest calibration error across nearly all groups. It is noteworthy that, while GCULR demonstrated superior accuracy in Table 2, its performance on the gASCE metric is not as impressive. This outcome aligns with theoretical expectations, as GCULR guarantees low group-conditional bias but does *not* guarantee calibration within each group, which is what we are measuring here.