UNSUPERVISED CONFORMAL INFERENCE: BOOTSTRAPPING AND ALIGNMENT TO CONTROL LLM UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying black-box LLMs requires managing uncertainty in the absence of token-level probability or true labels. We propose introducing an unsupervised conformal inference framework for generation, which integrates: generative models, incorporating: (i) an LLM-compatible atypical score derived from response-embedding Gram matrix, (ii) UCP combined with a bootstrapping variant (BB-UCP) that aggregates residuals to refine quantile precision while maintaining distribution-free, finite-sample coverage, and (iii) conformal alignment, which calibrates a single strictness parameter τ so a user predicate (e.g., factuality lift) holds on unseen batches with probability $\geq 1 - \alpha$. Across different benchmark datasets, our gates achieve close-to-nominal coverage and provide tighter, more stable thresholds than split UCP, while consistently reducing the severity of hallucination, outperforming lightweight per-response detectors with similar computational demands. The result is a label-free, API-compatible gate for test-time filtering that turns geometric signals into calibrated, goal-aligned decisions.

1 INTRODUCTION

Reliable *uncertainty quantification (UQ)* for large language models (LLMs) is needed for trustworthy AI. An assertive yet baseless claim can swiftly spread and cause damage, but for most practitioners, frontier models arrive only as black-box APIs with no access to gradients, exact log probabilities, or hidden states (Lin et al., 2024). Hence deployment teams must make keep-or-discard decisions from samples alone.

In black-box deployments, LLM uncertainty must be inferred from the sampled outputs themselves. Query-only signals include: (i) semantic-entropy methods that quantify dispersion across equivalence classes of responses and are effective for hallucination detection (Farquhar et al., 2024; Kossen et al., 2024); (ii) self-consistency, which uses agreement among independently sampled answers as a proxy for confidence (Wang et al., 2023; 2025); and (iii) geometry-based measures computed from response embeddings—e.g., local density or Gram-volume statistics—that correlate with quality and robustness (Qiu & Miikkulainen, 2024; Li et al., 2025). Because these signals require neither logits nor gradients, they are natural conformity scores for our unsupervised conformal calibration; in parallel, conformal wrappers for language modeling and factuality control are emerging (Quach et al., 2024; Mohri & Hashimoto, 2024).

Conformal prediction (CP) is model-agnostic and supplies finite-sample, distribution-free guarantees (Angelopoulos & Bates, 2022; Vovk et al., 2005b). However, the generative workflow breaks the classical supervised setting: prompts are not quantifiable covariates. A practical procedure must therefore calibrate in a *label-only* regime and use few, parallelizable API calls to expensive models.

We introduce a practical *unsupervised conformal prediction (UCP)* framework that increases data efficiency, reduces computation via bootstrapped conformal calibration, and reconciles heterogeneous modalities through conformal alignment, while delivering distribution-free, finite-sample guarantees and strong empirical gains in hallucination detection and factuality. Our framework calibrates directly on raw outputs and is compatible with black-box APIs:

1. an LLM-friendly *atypicality* score based on inner-product interaction energy of the response-embedding Gram matrix (unit-norm, cosine), yielding a bounded, exchangeability-compatible conformity score;
2. batched unsupervised conformal procedures—split UCP for single-batch queries and new batched variants (*B-UCP* and bootstrap-stabilized *BB-UCP*)—with finite-sample coverage under batch exchangeability and improved stability/efficiency over split UCP;
3. *conformal alignment*: a batch-level calibration of a single strictness knob τ that ensures any predicate (e.g., factuality improvement) holds on unseen batches with probability at least $1 - \alpha$, enabling label-free test-time gating

The structure of the paper is outlined as follows. §2 explores query-only UQ and the foundational unsupervised CP techniques. In §3, we introduce the concepts of Gram-matrix typicality, batched/bootstrapped calibration, and conformal alignment. Experimental outcomes are presented in §4, while §5 discusses conclusions and future research directions.

2 Background

2.1 Conformal Prediction for Generative Outputs

Conformal prediction (CP) is a model-agnostic, distribution-free method that turns arbitrary scores into set-valued inferences with finite-sample guarantees under a common exchangeability assumption (Vovk et al., 2005a; Angelopoulos & Bates, 2022; Lei et al., 2017). Classical assumptions can be relaxed via covariate-shift and dependence-aware extensions (Barber et al., 2023; Gibbs & Candès, 2021). Conformal Risk Control (Angelopoulos et al., 2025) not only addresses coverage but also aims to manage expected losses. Recent developments in language models use conformal calibration to enhance modeling accuracy, concentrating on removing and verifying claims (Quach et al., 2024; Mohri & Hashimoto, 2024; Cherian et al., 2024).

Building on the concepts in (Vovk et al., 2005a; Lei et al., 2017; Lei & Wasserman, 2012; Sadinle et al., 2018), we adopt the formal definitions of *Full-UCP* and *Split-UCP* as presented in Wasserman. The corresponding algorithms also appear in the Appendix. We observe the responses $Y_{1:n}$ paired with fixed prompts $X_{1:n}$ and a future pair (X_{n+1}, Y_{n+1}). We primarily calibrate on the responses Y (i.e., model outputs), treating prompts/contexts X as covariates that may be implicit. We treat $Y_{1:n}$ as exchangeable. We define residuals via a permutation-invariant map ϕ by $R_i = \phi(Y_i; \mathcal{S}_i)$. For tolerance $\alpha \in (0, 1)$, we target $\Pr Y_{n+1} \in C_n \geq 1 - \alpha$.

Full-UCP augments data with a candidate y_{n+1}, and residuals are recalculated for the augmented set $\{Y_1, \ldots, Y_n, y_{n+1}\}$. A conformal p-value is then derived from the residual ranking, incorporating y when $\pi(y) \geq \alpha$. **Split-UCP** partitions the data into sets $\mathcal{D}_1 = \{Y_i \colon i \in I_1\}$ and $\mathcal{D}_2 = \{Y_i \colon i \in I_2\}$, where index sets I_1 and I_2 partition the set $\{1, \ldots, n\}$. We then calculate calibration residuals $R_i = \phi(Y_i; \mathcal{D}_1)$ for $i \in I_2$ to establish the $(1-\alpha)(1+\frac{1}{|I_2|})$-quantile q of these residual values, subsequently yielding the set $C_n = \{y \colon \phi(y; \mathcal{D}_1) \leq q\}$. In addition, the Split-UCP marginal theorem states that under exchangeability of $(Y_{1:n}, Y_{n+1})$, the resulting set C_n satisfies $\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha$.

Full-UCP can be computationally inefficient due to retraining and ineffective searching over future candidates y (grid/root-finding). On the other hand, Split-UCP is data-inefficient as only the calibration split $\mathcal{D}_2$ influences the quantile, leading to sample-splitting costs. The challenges motivate us to design our framework for better alignment with the generative stochastic process $P(Y \mid X)$ to ensure that practical applications effectively capture underlying variability.

2.2 Gram Matrix Construction, Inner–Product Energy, and Atypical Score

To quantify LLM response uncertainty, we build our framework on the response-embedding Gram matrix. Given n responses $Y_1, \ldots, Y_n$ with embeddings $v_i := \psi(Y_i) \in \mathbb{R}^d$, we stack the embeddings as rows to create the matrix $V \in \mathbb{R}^{n \times d}$ with $V_{i,:} = v_i^\top$ and form the (uncentered) Gram $G := VV^\top \in \mathbb{R}^{n \times n}$ with entries $G_{ij} = \langle v_i, v_j \rangle$, where unit–norm embeddings $\|v_i\|_2 = 1$ by default.

We then define the inner–product (interaction) energy:

$$e(i; G) \;:=\; \|G_{:,i}\|_2 \;=\; \|V\, v_i\|_2. \tag{2.1}$$

In unit–norm embeddings, $e(i;G) = \left(\sum_{j=1}^{n} \cos^2 \theta_{ij}\right)^{1/2}$, where θ_{ij} denotes the angle between v_i and v_j. Thus, $e(i;G)^2$ quantifies the total squared directional alignment of v_i with its peers. Since $\cos^2$ treats aligned and anti-aligned directions equally, a *large* $e(i;G)$ indicates central, redundant elements (high agreement), while a *small* $e(i;G)$ denotes unique or irrelevant content (high novelty). The following theorem states that $e(\cdot;G)$ ranges from 1 to $\sqrt{n}$ (the proof is available in the Appendix).

Theorem 2.1 (Unit–norm interaction–energy bound)**.** *If $\|v_i\|_2 = 1$ for all i, then $1 \le e(i;G) \le \sqrt{n}$ for each i. Equality $e(i;G) = \sqrt{n}$ holds when v_i is perfectly aligned with all v_j; $e(i;G) = 1$ when v_i is orthogonal to all v_j for $j \neq i$.*

Letting B_E denote the supremum of $e(\cdot;G)$, we define the Atypical Score as

$$\Phi(i;G) \;:=\; 1 - \frac{e(i;G)}{B_E} \;\in [0,1]. \tag{2.2}$$

Under unit-norm embeddings, the upper bound B_E equals $\sqrt{n}$. Note that any strictly monotone transform of e is equivalent for ranking.

3 METHODOLOGY

3.1 BATCH UNSUPERVISED CONFORMAL PREDICTION

We adapt UCP to a *batched* setting that gathers information across exchangeable batches and may stabilize calibration using a within-batch bootstrap. For notational simplicity, assume there are $n+1 = (J+1)I$ responses $Y_1, \ldots, Y_{n+1}$ for pre-chosen integers J and I. Partition the $n+1$ responses into $J+1$ disjoint batches $\mathcal{B}_j = \{Y_{j,1}, \ldots, Y_{j,I}\}$ for $j = 1, \ldots, J+1$. The entire §3 assumes *batch exchangeability*, i.e., that the $J+1$ batches $\mathcal{B}_1, \ldots, \mathcal{B}_J,\ \mathcal{B}_{J+1}$ are i.i.d. and that within each batch the responses are exchangeable. (The latter condition is also called *partial exchangeability* (de Finetti, 1938).) Thus we will assume without loss of generality that $Y_{n+1} = Y_{J+1,I}$.

We design two statistically distinct methods under this setting:

- *Batch-UCP (B-UCP).* For each calibration batch j, residuals are computed within batches against leave-one-out Gram matrix base, where $R_{j,i} = \phi(Y_{j,i}; \mathcal{B}_{j,-i})$ using a bounded, permutation-invariant score $\phi \in [0, B_\phi]$. Pool $\{R_{j,i}\}_{j=1:J,\ i=1:I}$ and take a single adjusted conformal quantile.
- *Batch Bootstrap-UCP (BB-UCP).* For each calibration batch j, bootstrap the empirical residual multiset $\{R_{j,i}\}_{i=1}^{I}$ to obtain $\{S_{j,k}\}_{k=1}^{K}$. Pool $\{S_{j,k}\}_{j,k}$ and apply the same adjusted quantile. The bootstrap mitigates noise caused by irregularity from outlier batches, maintaining exchangeability.

The block below clarifies how B-UCP and BB-UCP differ. We present the formal algorithms in the Appendix (Algorithms B.3–B.4).

Unified Batch U-CP (demo; $K{=}0 \Rightarrow$ B-UCP, $K{\ge}1 \Rightarrow$ BB-UCP)

Inputs: $\{Y_k\}_{k=1}^{(J+1)I-1}$, score ϕ, batch count J, tolerance α, bootstrap count $K \ge 0$
Partition: calibration $\{\mathcal{B}_j\}_{j=1}^{J}$, hold-out $\mathcal{B}_{J+1,-I}$
for $j = 1$ **to** J **do**
  for $i = 1$ **to** I **do** $R_{j,i} \leftarrow \phi\big(Y_{j,i};\, \mathcal{B}_{j,-i}\big),\ \mathcal{B}_{j,-i} = \mathcal{B}_j \setminus \{Y_{j,i}\}$
  end for
  if $K > 0$ **then** draw $\{S_{j,k}\}_{k=1}^{K}$ from $\{R_{j,i}\}_{i=1}^{I}$
  end if
end for
Bag: $\mathcal{D} \leftarrow \{R_{j,i}\}_{j,i}$ if $K = 0$; else $\mathcal{D} \leftarrow \{S_{j,k}\}_{j,k}$
Quantile: $\delta_J = (J+1)\alpha - 1$; set $q \leftarrow B_\phi$ if $\delta_J \le 0$, else $q \leftarrow (1 - \delta_J/J)$-quantile of $\mathcal{D}$
Output: $C_n = \{y : \phi(y; \mathcal{B}_{J+1,-I}) \le q\}$

Under the batch exchangeability assumption, we have the following coverage guarantees, and we present the proofs in the Appendix.

Theorem 3.1 (B-UCP coverage)**.** *The prediction set C_n returned by Batch U-CP satisfies* $\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha$.

Theorem 3.2 (BB-UCP coverage)**.** *The prediction set C_n returned by Batch Bootstrap U-CP satisfies* $\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha$.

Our design incorporates three main mechanisms.

1) *Batching.* In the unsupervised setting, conventional CP loses the easy exchangeability of supervised CP because the conformity score depends on the other responses. By organizing data into exchangeable batches and using *within-batch leave-one-out* residuals (each $R_{j,i}$ computed against a base of size $I-1$), we effectively re-enable cross-validation–style conformalization.
2) *Within-batch LOO under batch exchangeability.* This alignment makes the calibration and test residual laws match, removing the split-sample penalty in split-UCP (which inflates the order-statistic index) and yielding tighter thresholds at a fixed risk level.
3) *Bootstrap aggregation.* Averaging replicated empirical laws within each batch stabilizes the empirical quantile and down-weights idiosyncratic batches, reducing the chance of underestimating the target quantile. Realized coverage therefore tends to be slightly conservative while intervals remain short. Resampling $\{R_{j,i}\}$ is inexpensive and preserves exchangeability; and because the method is rank-based, any strictly non-decreasing transform of ϕ leaves C_n unchanged. These effects anticipate our observations: BB-UCP is typically more conservative than split-UCP yet produces tighter, more stable intervals.

3.2 Conformal Alignment

Conformal alignment functions as a quality control technique across various modalities, enabling multilevel filtering and alignment superior to standard UCP schemes. Initially, we parameterize strictness using a single knob $\tau \in [0, 1]$ to filter batches with a low-cost, consistently available signal (here, the Gram matrix inner energy). During calibration, this signal is aligned with a rare or expensive quality measure (e.g., factuality), enabling deployment with just the low-cost signal while retaining the calibrated assurance $\Pr(\text{predicate on future batch}) \geq 1 - \alpha$. The idea scales to various contexts: establishing an accessible, cost-effective score allows the use of the same method to determine a global $\hat{\tau}$ from past batches, which is then applied to unlabeled new data. By utilizing text and Gram scores, the predicate can adapt any non-decreasing, right-continuous batch metric to inexpensive scores derived from text, vision, audio, or multimodal embeddings, offering substantial flexibility and a wide range of applications.

Similar to §3.1, partition the data into J disjoint batches $\mathcal{B}_j = \{Y_{j,1}, \ldots, Y_{j,I}\}$ for $j = 1, \ldots, J$, with $\{\mathcal{B}_j\}_{j=1}^{J+1}$ exchangeable and $\mathcal{B}_{J+1}$ the future batch. Let $\mathcal{RC}$ be the space of right-continuous, non-decreasing maps $[0, 1] \to [0, 1]$, and define $\psi : \binom{\mathcal{Y}}{I} \to \mathcal{RC}$. For each batch j, set $\mathcal{P}_j(\cdot) = \psi(\mathcal{B}_j)$; then $\mathcal{P}_j$ is non-decreasing and right-continuous, and $\{\mathcal{P}_j\}$ is exchangeable.

We use $\mathcal{P}_j(\tau)$ as a batch predicate with a subset-selection parameter $\tau \in [0, 1]$, which describes the j-th batch. For instance, let $\widehat{J}_j(\tau) \subseteq \{1, \ldots, I\}$ be right-continuous filtered sets; that is, for any $0 \leq \tau < \tau' \leq 1$ there exists $\delta > 0$ with $\widehat{J}_j(\tau') \subset \widehat{J}_j(\tau) = \widehat{J}_j(\tau + \delta)$. From the set $\widehat{J}_j(\tau)$ of indices, we define $\mathcal{P}_j(\tau)$ as the indicator of the event "$\widehat{J}_j(\tau)$ satisfies property A" where "property A" is to be determined according to a specific prediction target. We search for $\widehat{\tau}$ such that $\mathcal{P}_{J+1}(\widehat{\tau}) = 1$ with high probability, which means that the selected set $\widehat{J}_{J+1}(\hat{\tau})$ satisfies "property A" with high probability.

Define the minimal passing strictness

$$S_j \coloneqq \min\{\tau \in [0, 1] \colon \mathcal{P}_j(\tau) = 1\} \in [0, 1], \tag{3.1}$$

with $\inf \varnothing = 1$. Let $K = \lceil (1 - \alpha)(J + 1) \rceil$ and calibrate $\widehat{\tau}$ as the K-th order statistic of $\{S_j\}_{j=1}^{J}$.

Algorithm 3.1 Batch U-CP Conformal Alignment

1: **Input:** calibration batches $\mathcal{B}_1, \ldots, \mathcal{B}_J$; test batch $\mathcal{B}_{J+1,-I}$; $K \leftarrow \lceil (1-\alpha)(J+1) \rceil$; function ψ; tolerance $\alpha \in (0,1)$.
2: **for** $j = 1$ **to** J **do**
3: Compute $\mathcal{P}_j(\cdot) = \psi(\mathcal{B}_j)$
4: Compute $S_j = \inf\{\tau \in [0,1] \colon \mathcal{P}_j(\tau) = 1\}$.
5: **end for**
6: **Calibrate** $\widehat{\tau} \leftarrow$ the K-th smallest value among $\{S_j\}_{j=1}^{J}$. (If $K = J+1$, then $\hat{\tau} \leftarrow 1$) (Split-conformal quantile with $J+1$ total batches)
7: **Output** $\hat{\tau}$.

A remark of Algorithm 3.1 is that, for a target threshold $r \in (0,1)$ which we want $\mathcal{P}_{J+1}(\hat{\tau}) \geq r$ to hold for high probability, define $\mathcal{P}'_j(\tau) = \mathbf{1}\{\mathcal{P}_j(\tau) \geq r\}$. Then $\mathcal{P}'_j$ is also non-decreasing and right-continuous, and $\mathcal{P}_j(\tau) \geq r \iff \mathcal{P}'_j(\tau) = 1$. Hence Algorithm 3.1 records $\{S'_j\}$ and returns a single $\widehat{\tau}$ with $\Pr\{\mathcal{P}_{J+1}(\hat{\tau}) \geq r\} = \Pr\{\mathcal{P}'_{J+1}(\widehat{\tau}) = 1\} \geq 1 - \alpha$ under exchangeability.

Theorem 3.3 (B-UCP alignment guarantee)**.** *Assume the batches $\{B_j\}_{j=1}^{J+1}$ are exchangeable (which implies that predicates $\{\mathcal{P}_j\}_{j=1}^{J+1}$ are exchangeable) and that each $\mathcal{P}_j(\cdot)$ is non-decreasing and right-continuous in its argument τ for $j = 1, \ldots, J{+}1$. Then Algorithm 3.1 satisfies*

$$\Pr\big\{\ \mathcal{P}_{J+1}\big(\widehat{\tau}\big) = 1\ \big\}\ \geq\ 1 - \alpha.$$

By Theorem 3.3, any non-decreasing, right-continuous batch predicate calibrated across exchangeable batches yields a $1-\alpha$ guarantee on the held-out batch. We now instantiate this scheme for black-box LLMs by selecting a cheap Gram-geometry self-consistency score Q with the induced filter $\widehat{J}_j(\tau)$ and a batch predicate $\mathcal{P}_j$ that encodes deployment goals (e.g., factuality lift), as specified next.

Let $Q_{j,i}$ denote the inner-product interaction energy $e(i;G)$ from Section 2.2 (unit-norm, cosine geometry). We keep high-consensus items via

$$\widehat{J}_j(\tau)\ :=\ \{\, i :\ Q_{j,i} > \tau \,\},$$

so larger τ filters out more responses.

Let $s_{j,i} \in [0,1]$ be a batch severity with larger = worse (e.g., factuality severity). In strictness τ, set $K_j(\tau) = \{i : Q_{j,i} > \tau\}$ and $D_j(\tau) = \{i : Q_{j,i} \leq \tau\}$, and declare pass when the indicator is evaluated as 1, where

$$\mathcal{P}_j^{\mathrm{CVAR}}(\tau)\ :=\ \mathbf{1}\Big\{ \underbrace{\mathrm{CVAR}_q(s_{j,i} : i \in D_j(\tau)) - \mathrm{CVAR}_q(s_{j,i} : i \in K_j(\tau))}_{\Delta\mathrm{CVAR}_{j,\tau}(q)}\ \geq\ \delta \Big\}.$$

CVAR focuses on the worst tail. Requiring a positive gap means that the kept set reduces severe errors (rare but damaging hallucinations) relative to the dropped set (Chow et al., 2015; Zhao et al., 2025; Rockafellar & Uryasev, 2000; Acerbi & Tasche, 2002). Because $K_j(\tau)$ enlarges as τ decreases, $\tau \mapsto \mathcal{P}_j^{\mathrm{CVAR}}(\tau)$ is non-decreasing/right-continuous, so Alg. 3.1 applies unchanged.

Let $s_{j,i}$ be a factuality severity in $[0,1]$ (lower is better; e.g., BERTScore–F1 dissimilarity). The predicate $\mathcal{P}_j^{\mathrm{F}}(\tau) = 1$ asserts that the Q-filtered subset attains a statistically significant median reduction in factuality severity (per the test above). Calibrating $\hat{\tau}$ across historical batches yields a single label-free gate which, using Q alone at deployment, preserves this improvement on new batches with probability at least $1-\alpha$ (Theorem 3.3). *Applications include:* (i) open-domain QA/RAG: ensure only consistent answers are shown; (ii) customer support and search snippets: reduce the risk of false confident statements; (iii) summarization/reporting: exclude sections that do not pass factual accuracy checks before publishing.

4 Experiments

We perform three experiments to study three research questions aligned with §3: *(RQ1)* within a single query, does BB-UCP produce tighter prediction sets than S-UCP at the same target miscoverage?

(RQ2) across multiple queries, do B-/BB-UCP achieve nominal coverage and improve batch quality by discarding higher-severity responses? *(RQ3)* in cross-query alignment, can the calibrated global strictness $\hat{\tau}$ reliably yield batch-level severity reduction while preserving the conformal guarantee?

4.1 Experimental setup

We evaluate across four complementary QA datasets, each exhibiting a distinct failure mode: *ASQA*—ambiguity and underspecification (Stelmakh et al., 2023); *NQ-Open*—single-hop factoid retrieval (Lee et al., 2019; Kwiatkowski et al., 2019); *HotpotQA*—multi-hop composition (Yang et al., 2018); and *AmbigQA*—aliases and answer sets (Min et al., 2020). To probe sensitivity, we add two ablations: a decoding-entropy stress test and a vendor/model swap. For each open-domain QA prompt, we (i) synthesize a diverse response set by mixing plain answers, lightly *enforced* canonical answers, and structured *noise* outliers. This controlled injection is standard for stress test hallucination detection and semantic dispersion UQ signals (Kuhn et al., 2023; Qiu & Miikkulainen, 2024) and allows us to probe robustness under realistic contamination. All texts are embedded with a lightweight sentence encoder (`all-MiniLM-L6-v2`). We stack unit-normalized embedding vectors by rows to form V and then the Gram matrix $G = VV^\top$; (ii) expand the reference set with concise paraphrases to reduce aliasing; and (iii) characterize each candidate using a distance-based metric, **Factuality Severity** (FS). All artifacts are logged and kept provider-agnostic across OpenAI, Together, and Gemini (OpenAI et al., 2024; Grattafiori et al., 2024; Team et al., 2024).

We quantify answer quality with respect to references using BERTScore–F1 with baseline rescaling (`roberta-large`) (Zhang et al., 2020) on the *answer head* (first sentence or a `Final:` field, truncated to ≤ 16 tokens). Letting $\mathrm{head}(a)$ be the head of answer a and $\mathcal{R}_q$ the reference set, we define the severity

$$\mathrm{FS}(a) \;\coloneqq\; 1 - \max_{r \in \mathcal{R}_q} \mathrm{BERTScoreF1}\big(\mathrm{head}(a),\, r\big) \;\in [0,1], \tag{4.1}$$

so that a value of 0 indicates a near-paraphrase of some reference (high factual alignment), and a value near 1 flags semantic deviation. Scoring the head avoids rationale contamination and normalizes across style/length.

Given a batch B and the kept subset $K(\tau)$ after filtering by Q, we summarize factuality lifting by the *median reduction*

$$\Delta_{\mathrm{FS}}(\tau) = \mathrm{median}\{FS(y) : y \in B\} - \mathrm{median}\{FS(y) : y \in K(\tau)\}$$

(larger is better). All conformal calibrations use the Gram inner–energy score e from §3 and are implemented exactly as specified (*S-/B-/BB-UCP* and alignment). We fix random seeds, cache embeddings/Grams, and run identical pipelines on all datasets.

4.2 Experiment I — Single-Query Conformal Calibration (S-UCP vs. BB-UCP)

In the single-query regime, for each question q we embed all candidate responses, form the unit-norm response Gram matrix G, and compute the inner-product energy score Q for each response. We then construct residuals within the same pool and repeatedly split into calibration/test subsets of fixed sizes. Split UCP thresholds test residuals by directly taking a quantile of residuals; BB-UCP additionally bootstraps the calibration residuals within the batch and aggregates the quantiles to stabilize the threshold. We report (i) empirical coverage against the $1 - \alpha$ target and (ii) an efficiency proxy given by the accepted sublevel endpoint $q_{1-\alpha}$ (smaller is better).

Across *AmbigQA*, *AmbigQA-ENT*, *ASQA*, and *HotpotQA*, both S-UCP and BB-UCP achieve near-nominal coverage, while BB-UCP consistently achieves more conservative empirical coverage across repetitions and α and yields shorter interval length; see Fig. 1 (top/bottom). On *NQ-Open* and *NQ-Open-Vend*, performance is weaker: the answer pools are small and low-diversity, often collapsing into a single high-consensus "heap." This compresses dispersion in Gram space, produces near-tied residual ranks, and blunts the bootstrap's advantage; Split UCP 's coverage remains close to target. Slightly enlarging pool size or boosting response diversity effectively restores similar qualitative improvements as seen in other datasets.

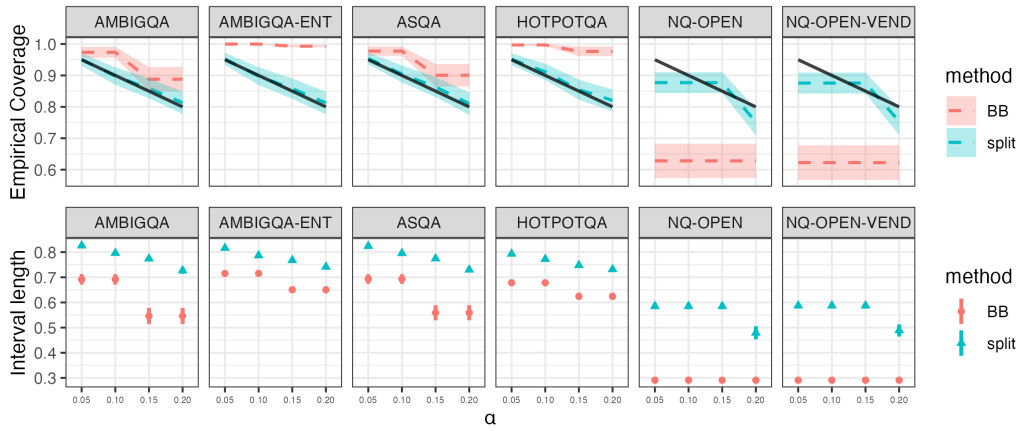

Figure 1: **Experiment 1. Top row:** Empirical coverage vs. α for a representative dataset (ASQA). Shaded bands show ± 1 SE, and the black line is the $1-\alpha$ target. Both the BB and Split UCP achieve the desired theoretical assurance, consistently staying above the target. BB is typically more conservative to attain greater empirical coverage. **Bottom row:** Interval size comparison bar are plotted across datasets, grouped by method (BB, split) and stratified by $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. In *all* datasets and at *all* α levels, BB consistently produces smaller interval sizes than Split, indicating the bootstrapping achieves statistical efficiency as intended.

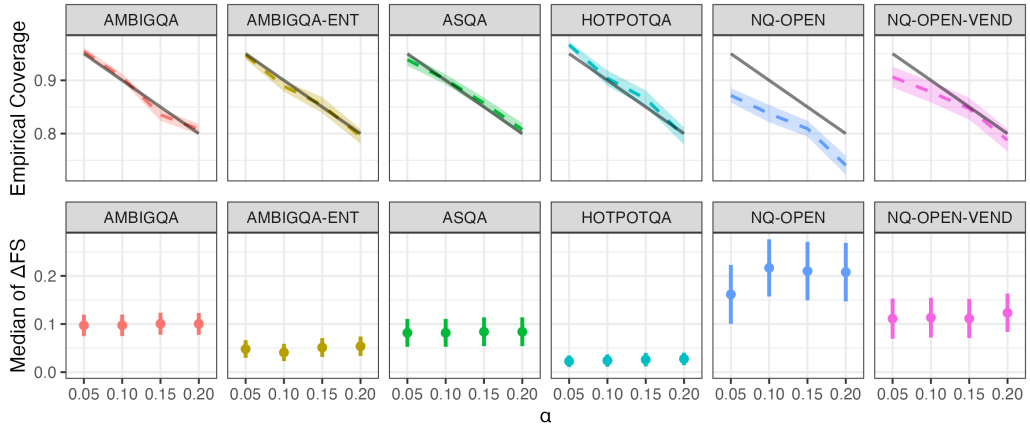

Figure 2: **Experiment 2. Top row:** Empirical coverage vs. α. Each panel corresponds to one dataset. Shaded bands show $\pm$ SE, and the black line is the $1-\alpha$ target. Across all datasets, cross-query BB-UCP nearly achieves the expected theoretical guarantee, with a single failure occurring solely in an extreme stress test. **Bottom row:** The change in median factuality severity (ΔFS) is shown across datasets (points indicate the mean, intervals represent $\pm$ SE). For every dataset, ΔFS constantly stay above 0, highlighting the efficacy of BB-UCP in improving factuality.

4.3 Experiment II — Cross-Query Calibration (BB-UCP) and Factuality Lifting

To mimic real-world deployment, each response set of the same query forms a *batch*. We run leave-one-query-out (LOQO) cross-validation: hold one batch for testing and calibrate on the rest. For every calibration batch we compute residuals from Q under within-batch LOO, bootstrap residuals per batch, pool the bootstrapped $(1-\alpha)$-quantiles into a single threshold $q^{\mathrm{BB}}_{1-\alpha}$, and apply that global threshold to the held-out batch. We evaluate (i) empirical coverage vs. $1-\alpha$ and (ii) factuality lifting via the median reduction in factuality severity, ΔFS (excluded minus kept; better positivity indicates improved prevention of hallucinations.).

LOQO empirical coverage closely tracks $1-\alpha$ across risk levels and datasets (Fig. 2, top), demonstrating that a *BB-UCP* global threshold learned from historical queries generalizes to unseen queries. More importantly for deployment, filtering by $q_{1-\alpha}^{\mathrm{BB}}$ consistently improves batch quality: $\Delta\mathrm{FS} > 0$ across *all* datasets and *all* α (Fig. 2, bottom; The hardest panels are instructive: on *NQ-Open*, median $\Delta\mathrm{FS} \approx 0.209$ (largest among benchmarks) despite average empirical coverage falling short of the 95% target (88.98% vs. 95.00%); *NQ-Open-Vend* shows the same trade-off (median $\Delta\mathrm{FS} \approx 0.112$, 92.88% vs. 95.00%) (Appendix C, Table 4). This under-coverage is attributable to small-N/low-entropy pools that yield discretization effects and near-ties in residual ranks. As in Experiment I, standard operational tweaks—increasing per-query pool size or adding response diversity—tighten coverage without erasing the observed factuality lift.

4.4 Experiment III — Cross-Query Conformal Alignment (CVaR-gap)

We perform LOQO folds for conformal alignment as well. In each fold, for every calibration batch j we scan a strictness grid τ and evaluate the CVaR-gap predicate $\mathcal{P}_j^{\mathrm{CVAR}}(\tau)$ at tail level q and margin δ, then record the minimal passing strictness S_j. We conformally calibrate a global $\hat{\tau}$ as the split-batch $(1-\alpha)$-quantile of $\{S_j\}$ (Alg. 3.1), then deploy on the held-out batch using only the cheap score Q: keep $K(\hat{\tau}) = \{i : Q_i > \hat{\tau}\}$ and report (i) empirical pass rate against $1-\alpha$ and (ii) factuality improvement via $\Delta\mathrm{CVAR}(\hat{\tau})$. Full predicate and implementation details are in Appendix B.5/C.

Alignment preserves the statistical target while delivering consistent factuality gains: for every dataset and every risk level, the reduction in factuality severity on the kept set is positive on average (Fig. 3). Notably, the largest median and mean improvements occur on the two hardest datasets—*NQ-Open* and *NQ-Open-Vend*—with median $\Delta\mathrm{FS} \approx 0.206$ and ≈ 0.112, respectively (Appendix C, Table 5). Aligning the affordable Gram-geometry score with the factuality signal provides an effective, economical filtering method, ensuring conformal assurance for new batches.

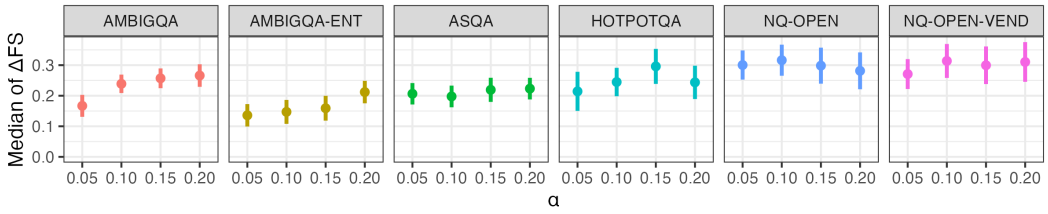

Figure 3: **Experiment 3.** Median of $\Delta\mathrm{FS}$ by dataset (points: mean, intervals: $\pm$ SE), and each panel corresponds to one dataset. Across *every* dataset and at *every* risk level, the reduction in factuality severity is consistently above 0 on average, demonstrating that conformal alignment with the gram matrix energy effectively enhances factuality while preserving statistical guarantees.

5 Conclusion

We introduced an unsupervised conformal inference framework for black-box LLMs that operates entirely on sampled outputs. The framework comprises three deployable components: (i) a Gram-geometry *atypicality* score based on unit-norm response-embedding inner products, yielding a bounded, interpretable, and stable signal; (ii) batched conformal procedures (B-UCP and the bootstrap-stabilized BB-UCP) that provide distribution-free, finite-sample guarantees under batch exchangeability while improving quantile stability and data efficiency over split UCP; and (iii) *conformal alignment*, which calibrates a global strictness parameter $\hat{\tau}$ so that a batch predicate (e.g., a CVaR-gap factuality lift) holds on new batches with probability at least $1-\alpha$. Conformal alignment provides a principled way to synchronize an expensive signal (ground truth) with a cheaper proxy and performs well when deployment relies only on the proxy, yielding a probabilistic approach to multimodal signal gating and filtering.

Limitations and outlook. Our guarantees assume exchangeability across batches and within-batch permutation invariance; violations due to drift or covariate shift motivate weighted or

covariate-aware variants. Performance depends on embedding quality and normalization, underscoring the need for robustness audits and principled model/embedding selection. Predicate design (e.g., CVaR vs. median lifts and multi-metric trade-offs) invites cost-aware utilities and multi-task calibration. Extending alignment to multimodal settings and adding adaptive or online recalibration are promising directions for stronger reliability under non-stationarity.

Use of AI for language editing. We used OpenAI ChatGPT and Overleaf Writefull solely for language polishing (grammar, clarity, style) of author-written text. All ideas, experiments, and conclusions are the authors' own, and the authors reviewed and take responsibility for all content.

REFERENCES

Carlo Acerbi and Dirk Tasche. Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes*, 31(2):379–388, 2002.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2025. URL https://arxiv.org/abs/2208.02814.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2023. URL https://arxiv.org/abs/2202.13415.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods, 2024. URL https://arxiv.org/abs/2406.09714.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach, 2015. URL https://arxiv.org/abs/1506.02188.

Bruno de Finetti. Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, (739):5–18, 1938.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, jun 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift, 2021. URL https://arxiv.org/abs/2106.00170.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid

El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,

Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering, 2019. URL https://arxiv.org/abs/1906.00300.

Jing Lei and Larry Wasserman. Distribution free prediction bands, 2012. URL https://arxiv.org/abs/1203.5422.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017. URL https://arxiv.org/abs/1604.04173.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms, 2025. URL https://arxiv.org/abs/2502.21239.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL https://openreview.net/forum?id=XJiN1VkgA0.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions, 2020. URL https://arxiv.org/abs/2004.10645.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees, 2024. URL https://arxiv.org/abs/2402.10978.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian,

Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi,

Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space, 2024. URL https://arxiv.org/abs/2405.13845.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2024. URL https://arxiv.org/abs/2306.10193.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, June 2018. ISSN 1537-274X. doi: 10.1080/01621459.2017.1395341. URL http://dx.doi.org/10.1080/01621459.2017.1395341.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. Asqa: Factoid questions meet long-form answers, 2023. URL https://arxiv.org/abs/2204.06092.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine,

Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent

Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon

Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005a. ISBN 978-0387001524.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005b. ISBN 978-0-387-00152-4.

Tianyu Wang, Akira Horiguchi, Lingyou Pang, and Carey E. Priebe. Llm web dynamics: Tracing model collapse in a network of llms, 2025. URL https://arxiv.org/abs/2506.15690.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Larry Wasserman. Conformal prediction. Lecture notes for Statistical Machine Learning, Carnegie Mellon University. https://www.stat.cmu.edu/~larry/=sml/Conformal.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Yujie Zhao, Jose Efraim Aguilar Escamill, Weyl Lu, and Huazheng Wang. Ra-pbrl: Provably efficient risk-aware preference-based reinforcement learning, 2025. URL https://arxiv.org/abs/2410.23569.

A PROOFS

A.1 GRAM MATRIX

Theorem 2.1 (Unit–norm interaction–energy bound). *If $\|v_i\|_2 = 1$ for all i, then $1 \le e(i;G) \le \sqrt{n}$ for each i. Equality $e(i;G) = \sqrt{n}$ holds when v_i is perfectly aligned with all v_j; $e(i;G) = 1$ when v_i is orthogonal to all v_j for $j \neq i$.*

Proof.

$$e(i;G) := \|G_{:,i}\|_2 = \Big(\sum_{j=1}^{n} \langle v_i, v_j\rangle^2\Big)^{1/2} = \Big(\|v_i\|^2 + \sum_{j\neq i}^{n} \langle v_i, v_j\rangle^2\Big)^{1/2}$$

Since $\|v_i\|_2 = 1$, the $j = i$ term equals 1, so

$$e(i;G)^2 = 1 + \sum_{j\neq i} \langle v_i, v_j\rangle^2.$$

Now for $j \neq i$, because both v_i and v_j are unit vectors, $\langle v_i, v_j\rangle = \cos\theta_{ij} \in [0,1]$. Therefore,

$$1 \le e(i;G)^2 \le 1 + (n-1) = n,$$

and taking square roots gives the bound

$$1 \;\le\; e(i;G) \;\le\; \sqrt{n}.$$

The lower bound is attained when v_i is orthogonal to all v_j's with $j \neq i$; the upper bound is attained when v_i is perfectly aligned or anti-aligned with all v_j's. □

A.2 SPLIT UCP

Theorem A.1 (Split-UCP marginal coverage). *For exchangeable $\{Y_i\}_{i=1}^{n+1}$ and ϕ, the split conformal prediction set as stated in Section 3 satisfies $\Pr\{Y_{n+1} \in C_n\} \ge 1-\alpha$.*

Proof. Let $R_{n+1} = \phi(Y_{n+1}; \mathcal{D}_1)$. Given $\mathcal{D}_1$, the distribution of $(R_i)_{i\in I_2\cup\{n+1\}}$ is exchangeable. Hence

$$\begin{aligned}
\Pr\{Y_{n+1} \in C_n\} &= \Pr\Big\{R_{n+1} \;\le\; \text{ the } \lceil (1-\alpha)(|I_2|+1)\rceil\text{-th smallest of } \{R_i\}_{i\in I_2}\Big\} \\
&= \Pr\Big\{R_{n+1} \;\le\; \text{ the } \lceil (1-\alpha)(|I_2|+1)\rceil\text{-th smallest of } \{R_i\}_{i\in I_2\cup\{n+1\}}\Big\} \\
&= \frac{\lceil (1-\alpha)(|I_2|+1)\rceil}{|I_2|+1} \;\ge\; 1-\alpha.
\end{aligned}$$

□

A.3 BATCH U-CP

Theorem 3.1 (B-UCP coverage). *The prediction set C_n returned by Batch U-CP satisfies $\Pr\{Y_{n+1} \in C_n\} \ge 1-\alpha$.*

Proof. If $\frac{1}{J+1} \ge \alpha$, then C_n is the whole space and the claim is trivial. Now assume $\frac{1}{J+1} < \alpha$. Define the (random) indicator function

$$L_{j,i}(\lambda) := \mathbf{1}\big\{\phi(Y_{j,i}, \mathcal{B}_{j,-i}) > \lambda\big\}, \qquad j = 1,\dots,J{+}1,\;\; i = 1,\dots,I.$$

Since $Y_{J+1,I}$ is not used to construct C_n, we have

$$\Pr(Y_{n+1} \in C_n) = \Pr\big(\phi(Y_{n+1}, \mathcal{B}_{J+1,-I}) \le q\big) = 1 - \mathbb{E}[L_{J+1,I}(q)]\,,$$

where q is the $\big(1 - \frac{(J+1)\alpha - 1}{J}\big)$-quantile of the set $\{\phi(Y_{j,i}, \mathcal{B}_{j,-i})\}_{j=1:J,\, i=1:I}$. Define

$$\hat\lambda' := \inf\left\{\lambda\colon \tfrac{1}{(J+1)I}\sum_{j=1}^{J+1}\sum_{i=1}^{I} L_{j,i}(\lambda) \le \alpha\right\}, \quad \hat\lambda := \inf\left\{\lambda\colon \tfrac{1}{(J+1)I}\sum_{j=1}^{J}\sum_{i=1}^{I} L_{j,i}(\lambda) + \tfrac{1}{J+1} \le \alpha\right\}.$$

These two λ's exist because $L_{j,i}(B_\phi) = 0$ for any (i, j).

Then $\hat{\lambda}' \le \hat{\lambda} \le q$, so $L_{J+1,i}(\hat{\lambda}') \ge L_{J+1,i}(\hat{\lambda}) \ge L_{J+1,i}(q)$ for all i. Therefore

$$\mathbb{E}[L_{J+1,I}(q)] \le \mathbb{E}\left[L_{J+1,I}(\hat{\lambda})\right] = \mathbb{E}\Big[\mathbb{E}\Big[L_{J+1,I}(\hat{\lambda}) \;\Big|\; \{\mathcal{B}_j\}_{j=1}^{J}\Big]\Big] = \mathbb{E}\left[\tfrac{1}{I}\sum_{i=1}^{I} L_{J+1,i}(\hat{\lambda})\right]$$

where the last equality uses exchangeability within batch B_{J+1}. Because $\hat{\lambda}' \le \hat{\lambda}$, the right-most term is bounded above by

$$\mathbb{E}\left[\tfrac{1}{I}\sum_{i=1}^{I} L_{J+1,i}(\hat{\lambda}')\right] = \mathbb{E}\left[\tfrac{1}{(J+1)I}\sum_{j=1}^{J+1}\sum_{i=1}^{I} L_{j,i}(\hat{\lambda}')\right] \le \mathbb{E}\left[\tfrac{1}{(J+1)I}\sum_{j=1}^{J+1}\sum_{i=1}^{I} L_{j,i}(\hat{\lambda}')\right] \le \alpha$$

by also using exchangeability of the batches $\{\mathcal{B}_j\}_{j=1}^{J+1}$, the inequality $\hat{\lambda} \le q$, and then the definition of q. Thus $\Pr(Y_{n+1} \in C_n) \ge 1 - \alpha$. □

A.4 Batch Bootstrap U-CP

Theorem 3.2 (BB-UCP coverage)**.** *The prediction set C_n returned by Batch Bootstrap U-CP satisfies* $\Pr\{Y_{n+1} \in C_n\} \ge 1 - \alpha$.

Proof. As with the proof in Section A.3, we have

$$\Pr\big(Y_{n+1} \in C_n\big) = \Pr\big(\phi(Y_{n+1}, \mathcal{B}_{J+1,-I}) \le q\big) = 1 - \mathbb{E}\big[\mathbf{1}\{\phi(Y_{n+1}, \mathcal{B}_{J+1,-I}) > q\}\big].$$

Also do virtual bootstrap in the future batch $\mathcal{B}_{J+1}$ to get $\{S_{J+1,k}\}_{k=1}^{K}$. Independence across j and identical distributions imply

$$\mathbb{E}\big[\mathbf{1}\{\phi(Y_{n+1}, \mathcal{B}_{J+1,-I}) > q\}\big] = \mathbb{E}\left[\tfrac{1}{K}\sum_{k=1}^{K} \mathbf{1}\{S_{J+1,k} > q\}\right].$$

Let

$$\hat{\lambda}' := \inf\left\{\lambda : \frac{1}{(J+1)K}\sum_{j=1}^{J+1}\sum_{k=1}^{K} \mathbf{1}\{S_{j,k} > \lambda\} \le \alpha\right\}$$

$$\hat{\lambda} := B_\phi \wedge \inf\left\{\lambda : \frac{1}{(J+1)K}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{1}\{S_{j,k} > \lambda\} + \frac{1}{J+1} \le \alpha\right\}.$$

Then $\hat{\lambda}' \le \hat{\lambda} \le q$ and the same exchangeability argument yields

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}\{\phi(Y_{n+1}, \mathcal{B}_{J+1,-I}) > q\}\big] &= \mathbb{E}\left[\tfrac{1}{K}\sum_{k=1}^{K} \mathbf{1}\{S_{J+1,k} > q\}\right] \\
&\le \mathbb{E}\left[\tfrac{1}{K}\sum_{k=1}^{K} \mathbf{1}\{S_{J+1,k} > \hat{\lambda}'\}\right] \\
&= \mathbb{E}\left[\tfrac{1}{(J+1)K}\sum_{j=1}^{J+1}\sum_{k=1}^{K} \mathbf{1}\{S_{j,k} > \hat{\lambda}'\}\right] \\
&\le \alpha,
\end{aligned}$$

hence $\Pr(Y_{n+1} \in C_n) \ge 1 - \alpha$. □

A.5 Batch-wise Conformal Alignment

Theorem 3.3 (B-UCP alignment guarantee)**.** *Assume the batches $\{B_j\}_{j=1}^{J+1}$ are exchangeable (which implies that predicates $\{\mathcal{P}_j\}_{j=1}^{J+1}$ are exchangeable) and that each $\mathcal{P}_j(\cdot)$ is non-decreasing and right-continuous in its argument τ for $j = 1, \ldots, J+1$. Then Algorithm 3.1 satisfies*

$$\Pr\{\ \mathcal{P}_{J+1}(\widehat{\tau}) = 1\ \}\ \geq\ 1 - \alpha.$$

Proof. The algorithm statement provides strictness $S_j := \inf\{\tau : \mathcal{P}_j(\tau) = 1\}$ with $\inf \varnothing = 1$, value $K = \lceil (1-\alpha)(J+1) \rceil$, and $\hat{\tau}$ being the K-th order statistic of $\{S_j\}_{j=1}^{J}$. Let $\hat{\tau}_{J+1}$ be the K-th order statistic of $\{S_j\}_{j=1}^{J+1}$. Noting that exchangeability of $\{\mathcal{P}_j\}_{j=1}^{J+1}$ implies exchangeability of $\{S_j\}_{j=1}^{J+1}$, we get

$$\Pr\{\mathcal{P}_{J+1}(\hat{\tau}) = 1\} = \Pr\{S_{J+1} \leq \hat{\tau}\} = \Pr\{S_{J+1} \leq \hat{\tau}_{J+1}\} = \frac{K}{J+1} \geq 1 - \alpha.$$

□

Remark. The proof skills for Theorem 3.1 and 3.2 comes from the attempt to go back from conformal risk control to conformal prediction. In the proof we constructed a proper loss function to achieve this with a standard argument of CRC. This technique can be transfer to many other settings where we need to go back to CP from CRC.

B Algorithms

B.1 Classical Full UCP (background)

Algorithm B.1 Full Unsupervised Conformal Prediction (Full-UCP)

1: **Input:** Data $Y_{1:n}$, score function ϕ, tolerance $\alpha \in (0, 1)$
2: **Output:** prediction set C_n
3: **for candidate** $y \in \mathbb{R}$ (grid or root-finding) **do**
4: Form $\mathcal{A} \leftarrow \{Y_1, \ldots, Y_n, y\}$
5: Compute residuals $R_i \leftarrow \phi(Y_i; \mathcal{S}_i)$ for $i = 1, \ldots, n$, and $R_{n+1} \leftarrow \phi(y; \mathcal{S}_{n+1})$
6: Compute the p-value $\pi(y) \leftarrow \dfrac{1}{n+1} \displaystyle\sum_{i=1}^{n+1} \mathbf{1}\{R_i \geq R_{n+1}\}$
7: **end for**
8: **Return** $C_n \leftarrow \{y : \pi(y) \geq \alpha\}$

B.2 Split UCP (background)

Algorithm B.2 Split Unsupervised Conformal Prediction (Split-U-CP)

1: **Input:** $Y_{1:n}$, score ϕ, tolerance α
2: Randomly form I_1, I_2; define $\mathcal{D}_1$
3: Compute residuals $R_i \leftarrow \phi(Y_i; \mathcal{D}_1)$ for $i \in I_2$
4: $q \leftarrow \mathcal{Q}_{(1-\alpha)(1+\frac{1}{|I_2|})}(\{R_i : i \in I_2\})$
5: **Return** $C_n \leftarrow \{y : \phi(y; \mathcal{D}_1) \leq q\}$.

B.3 BATCH UCP AND BATCH BOOTSTRAP U-CP (FORMAL ALGORITHMS)

Algorithm B.3 Batch U-CP

Input: responses $\{Y_k\}_{k=1}^{(J+1)I-1}$, score ϕ (bounded by B_ϕ), batch count J, tolerance α.
Partition the data into $J+1$ disjoint batches $B_j = \{Y_{j,1}, \dots, Y_{j,I}\}$ for $j = 1, \dots, J$ and $B_{J+1,-I} = \{Y_{J+1,1}, \dots, Y_{J+1,I-1}\}$.
for $j = 1$ **to** J **do** ▷ Within-batch leave-one-out residuals
  for $i = 1$ **to** I **do**
    $R_{j,i} \leftarrow \phi\big(Y_{j,i}; B_{j,-i}\big)$ where $B_{j,-i} = B_j \setminus \{Y_{j,i}\}$
  end for
end for
if $\frac{1}{J+1} \le \alpha$ **then**
  $q \leftarrow \big(1 - \frac{(J+1)\alpha - 1}{J}\big)$-quantile of $\{R_{j,i}\}_{j,i}$
else
  $q \leftarrow B_\phi$
end if
Return $C_n = \big\{y : \phi\big(y; B_{J+1,-I}\big) \le q\big\}$

Algorithm B.4 Batch Bootstrap U-CP

Input: responses $\{Y_k\}_{k=1}^{(J+1)I-1}$, score ϕ (bounded by B_ϕ), batch count J, tolerance α, bootstrap count K.
Partition as in Algorithm B.3.
for $j = 1$ **to** J **do** ▷ Within-batch leave-one-out residuals
  for $i = 1$ **to** I **do**
    $R_{j,i} \leftarrow \phi\big(Y_{j,i}; B_{j,-i}\big)$
  end for
  Draw K bootstrap replicates $\{S_{j,\ell}\}_{\ell=1}^{K}$ from $\{R_{j,i}\}_{i=1}^{I}$
end for
if $\frac{1}{J+1} \le \alpha$ **then**
  $q \leftarrow \big(1 - \frac{(J+1)\alpha - 1}{J}\big)$-quantile of $\{S_{j,\ell}\}_{j,\ell}$
else
  $q \leftarrow B_\phi$
end if
Return $C_n = \big\{y : \phi\big(y; B_{J+1,-I}\big) \le q\big\}$

B.4 ALGORITHMIC DETAILS FOR THE CVAR-GAP PREDICATE

Let $Q_{j,i}$ denote the inner-product interaction energy from Section 2.2 (unit-norm, cosine geometry). We keep high-consensus items via

$$\widehat{J}_j(\tau) \;:=\; \{\, i :\; Q_{j,i} > \tau \,\},$$

so larger τ retains fewer and more self-consistent responses.

Let $s_{j,i} \in [0,1]$ be a batch severity with larger = worse (e.g., factuality severity). At strictness τ, define the kept and dropped sets $K_j(\tau) = \{i : Q_{j,i} > \tau\}$ and $D_j(\tau) = \{i : Q_{j,i} \le \tau\}$. For a random variable X with CDF F_X, the upper-tail *Conditional Value-at-Risk* at level $q \in (0,1)$ is

$$\mathrm{CVAR}_q(X) \;:=\; \frac{1}{1-q}\int_q^1 \mathrm{VaR}_u(X)\, du, \quad \text{where } \mathrm{VaR}_u(X) = \inf\{x :\; F_X(x) \ge u\}.$$

We instantiate a batch predicate that asks for a *tail-risk improvement* after filtering:

$$\Delta\mathrm{CVAR}_{j,\tau}(q) \;:=\; \mathrm{CVAR}_q\big(s_{j,i} : i \in D_j(\tau)\big) \;-\; \mathrm{CVAR}_q\big(s_{j,i} : i \in K_j(\tau)\big),$$
$$\mathcal{P}_j^{\mathrm{CVAR}}(\tau) \;:=\; \mathbf{1}\{\Delta\mathrm{CVAR}_{j,\tau}(q) \ge \delta\}.$$

ID	Benchmark (split)	#Q	Para P	Ans N	Mix (N/E/Z)	Entropy τ
C1	ASQA (dev)	60	10	150	(.75/.00/.25)	0.90
C2	NQ-Open (val)	60	6	16	(.67/.00/.33)	0.86
C3	HotpotQA (val)	60	10	100	(.60/.00/.40)	0.86
C4	AmbigQA (dev)	60	10	150	(.75/.00/.25)	0.86
C5	AmbigQA (dev) (ablation: decoding entropy)	40	10	150	(.75/.00/.25)	0.86
C6	NQ-Open (val) (ablation: vendor/model)	60	6	16	(.67/.00/.33)	0.86

Table 1: Benchmarks and per-item sampling settings used in the hallucination study. The mix column shows (normal/enforced/noise).

CVAR, also known as *Expected Shortfall*, focuses on the worst tail and is coherent and robust to heavy tails; demanding a positive CVAR gap concentrates the kept set on reliably low-severity answers and suppresses rare but severe failures.

Algorithm B.5 CVaR-gap alignment: minimal strictness and split-batch calibration

1: **Inputs:** calibration batches $\{B_j\}_{j=1}^{J}$; held-out test batch B_{J+1}; Gram score $Q \in [0,1]$; tail level $q \in (0,1)$; margin $\delta \geq 0$; miscoverage $\alpha \in (0,1)$.
2: **Kept/excluded at strictness τ:** $K_j(\tau) = \{i : Q_{j,i} > \tau\},\ E_j(\tau) = B_j \setminus K_j(\tau)$.
3: **Empirical** CVaR_q**.** For a multiset $S \subset [0,1]$ with $m = |S|$, sort in descending order $s_{(1)} \geq \cdots \geq s_{(m)}$ and set $h = \lceil (1-q)m \rceil$, $\widehat{\mathrm{CVaR}}_q(S) = \frac{1}{h}\sum_{\ell=1}^{h} s_{(\ell)}$ (winsorize if h=0).
4: **Predicate:** $\mathcal{P}_j^{\mathrm{CVaR}}(\tau) = \mathbf{1}\big\{ \widehat{\mathrm{CVaR}}_q(s|E_j(\tau)) - \widehat{\mathrm{CVaR}}_q(s|K_j(\tau)) \ \geq\ \delta \ \big\}$.
5: **Minimal strictness (per batch).** Scan τ over the right-continuous grid induced by the unique Q values in B_j (plus 0 and 1). Let $S_j = \inf\{\tau \in [0,1] : \mathcal{P}_j^{\mathrm{CVaR}}(\tau) = 1\}$ (set $S_j \leftarrow 1$ if the set is empty).
6: **Split-batch calibration.** Return $\widehat{\tau} = Quant_{1-\alpha}(\{S_j\}_{j=1}^{J})$ as in §3.2.
7: **Deployment on** B_{J+1}**.** Keep $K_{J+1}(\widehat{\tau}) = \{i : Q_{J+1,i} > \widehat{\tau}\}$ and report the CVaR-gap.

Let $FS_{j,i}$ be a factuality severity in $[0,1]$ (lower is better; e.g., BERTScore–F1 dissimilarity). The predicate $\mathcal{P}_j^{\mathrm{F}}(\tau) = 1$ asserts that the Q-filtered subset achieves a statistically significant median reduction in factuality severity (per the test above). Calibrating $\hat{\tau}$ across historical batches yields a single label-free gate which, when applied with Q alone, preserves this improvement on new batches with probability at least $1 - \alpha$ (Theorem 3.3).

C Experiment

C.1 Appendix: Hallucination Experiment Settings and Configurations

For each question we generate a response set, compute **Factuality Severity** $= 1 - \max_{r \in \text{refs}} \mathrm{BERTScore}\text{-}F1(a, r)$. All runs are seeded and logged to timestamped, self-describing CSVs: a per-*answer* file (scores, margins, types, decoding knobs) and a per-*run* file (dataset/split, sample counts, model/provider, seeds, thresholds, and paths). Together model IDs are normalized to serverless fallbacks to avoid availability regressions.

We evaluate across four core datasets—ASQA (dev), NQ-Open (validation), HotpotQA (validation), AmbigQA (dev)—plus two ablations that stress decoding entropy and vendor/model choice. Each configuration fixes decoding knobs and the normal/enforced/noise mix, while paraphrasing a canonical gold to reduce aliasing of surface forms.

Shared knobs: alias-normalization for Together; `n_per_call`=5; rate-limit $\approx$ 0.8s; severity mix weight logged; seeds: C1=42, C2=7, C3=11, C4=23, C5=23, C6=8.

We use minimal, auditable prompts. For *paraphrasing* the canonical gold: *System:* "You rewrite text. Output a succinct standalone paraphrase." *User:* "Paraphrase the following answer in different wording, preserving the exact meaning and factual content. Keep it concise and standalone. Avoid hedging, qualifiers, or extra details. **Answer:** {*gold*}." For *normal answers*: *System:* "Answer the question with the canonical short answer

ID	Provider	Model	Temp	Top-p	MaxTok	Embed	BERTScore
C1	Together	Llama-3.3-70B-Instr. Turbo	1.3	1.0	256	MiniLM-L6-v2	RoBERTa-large
C2	OpenAI	gpt-4o-mini	0.1	1.0	96	MiniLM-L6-v2	RoBERTa-large
C3	Together	Mixtral-8x7B-Instr. v0.3	1.2	1.0	256	MiniLM-L6-v2	RoBERTa-large
C4	Together	Llama-3.1-8B-Instr. Turbo	0.7	0.9	256	MiniLM-L6-v2	RoBERTa-large
C5	Together	Llama-3.1-8B-Instr. Turbo	1.3	1.0	256	MiniLM-L6-v2	RoBERTa-large
C6	Together	Llama-3.1-8B-Instr. Turbo	0.1	1.0	96	MiniLM-L6-v2	RoBERTa-large

Table 2: Provider/decoding and measurement settings, linked by **ID** to Table 1.

first; then add at most one brief justification. Be concise." *User:* {*question*}. For *enforced canonical answers*: *System:* "Answer with the canonical short answer first; then a single, concrete supporting detail. Avoid aliasing, avoid hedging, avoid contradictory statements." *User:* {*question*}. (Noise/outlier strings are programmatically injected: gibberish, off-topic, fabricated citations, prompt-injection strings, contradictions, emoji floods, and multilingual snippets.)

Embeddings: `sentence-transformers/all-MiniLM-L6-v2` with unit-norm rows; semantic-entropy uses a soft neighbor kernel above τ (exponent κ=4) and a normalized $-\log$ mapping to $[0, 1]$. Severity-F1 uses `bert-score` with `roberta-large` (baseline-rescaled) on "answer heads" (first $\leq$16 tokens) to limit verbosity bias. All artifacts are timestamped and saved as {`dataset`}_{`model`}_{`stamp`}__`ns`{`N`} for direct reuse in downstream risk control. This mirrors the same compute-aware calibration-to-deployment recipe we use for LLM-as-Judge.

C.2 More Results

Table 3: **Benchmark mapping used in the six-panel comparisons for plotting.** Short codes are the compact labels used in figure titles. For JudgeQ CSVs, the same names appear with the suffix `__judged`.

Panel	Short code	CSV `dataset_name`
1	AMBIGQA-ENT	`ambigqa__llama8b__hiT__ablation_entropy__ns40_responses`
2	AMBIGQA	`ambigqa__llama8b__midT__ns60_responses`
3	ASQA	`asqa__llama70b__hiT__ns60_responses`
4	HOTPOTQA	`hotpot__mixtral8x7b__hiT_noise40__ns60_responses`
5	NQ-OPEN	`nq__gpt4omini__loT_light__ns60_responses`
6	NQ-OPEN-VEND	`nq__llama8b__loT__ablation_vendor__ns60_responses`

Across Experiment 1, our methods perform strongly on the majority of benchmarks: on ASQA, HOTPOTQA, and both AMBIGQA panels (standard and entropy-ablation), Split-UCP consistently achieves nominal coverage while BB-UCP further tightens the coverage range and reduces variability—demonstrating the intended stability benefit of within-batch bagging under heterogeneous answer clouds with heavier tails. These results highlight two key strengths of our approach: (i) *distribution-free coverage* remains intact across diverse datasets, providers, and decoding knobs, and (ii) *practical efficiency* improves in harder settings, where bootstrapping stabilizes tail quantiles and yields shorter, more reliable acceptance regions at a fixed risk level.

NQ-OPEN and its variant NQ-OPEN-VEND show weakness due to small query pools and low-diversity factoids, leading to compressed Gram-space dispersion. This results in nearly tied residual ranks, reducing the visibility of BB-UCP's advantage over Split-UCP.This behavior is an artifact of the small-N/low-entropy regime rather than a failure of validity, and is readily mitigated in practice by slightly increasing the pool size or calibration-only diversity (or by modest geometry smoothing), after which these panels recover the same qualitative gains observed on ASQA, HOTPOTQA, and AMBIGQA.

Experiment 2 Experiment 2 shows that the factuality lift is not only consistent but *strongest on the hardest panels*. Aggregating the raw bars across all α values, the median ΔFS is *strictly positive for every dataset* (all panels, all α), confirming that the Q-gate reliably reduces factuality severity on the kept set. Moreover, the largest median gains occur on NQ-OPEN and NQ-OPEN-VEND (*aka* EnqueueOpen/EnqueueOpenVend): NQ-OPEN achieves the top median improvement ($\tilde{\Delta}\mathrm{FS} \approx 0.209$), with NQ-OPEN-VEND second ($\tilde{\Delta}\mathrm{FS} \approx 0.112$), while the remaining benchmarks (ASQA, HOTPOTQA, AMBIGQA, AMBIGQA-ENT) are all positive as well. This "worst-case best" pattern indicates our gate concentrates probability mass on the most reliable answers precisely where the answer cloud is small and low-diversity. Coverage is slightly under nom-

Table 4: Experiment 2 aggregated results across miscoverage levels. ΔFS is the median-factuality reduction (excluded − kept); positive is better. Avg Cov. is empirical coverage averaged over α, Target Cov. is the average nominal $1-\alpha$, and Avg Gap is the mean (coverage − target) in percentage points (pp).

Benchmark	#α	Median ΔFS	Mean ΔFS	Avg Cov. (%)	Target Cov. (%)	Avg Gap (pp)
NQ-OPEN	5	0.209	0.189	88.98	95.00	-6.02
NQ-OPEN-VEND	5	0.112	0.101	92.88	95.00	-2.12
ASQA	5	0.091	0.089	95.53	95.00	0.53
AMBIGQA	5	0.072	0.073	95.55	95.00	0.55
HOTPOTQA	5	0.067	0.067	95.11	95.00	0.11
AMBIGQA-ENT	5	0.051	0.052	95.79	87.50	0.79

inal on average in those two hardest panels (mean gap ≈ -6.0 pp on NQ-OPEN; range of mean gaps across benchmarks $\approx [-6.0, +0.8]$ pp), which is expected from small-N discretization and near-tied ranks; it is also *actionable*—increasing the per-query pool or adding calibration-only diversity closes the shortfall without altering the factuality lift. Net: Experiment 2 provides a strong, data-backed claim of *robustness* (positive lift everywhere), *effectiveness* (largest gains on the hardest datasets), and *practical tunability* (coverage can be tightened by modest, standard knobs).

Table 5: Experiment 3 aggregated factuality reductions across miscoverage levels. ΔFS is the reduction in factuality severity (excluded − kept); positive is better. Columns summarize the distribution of per-α improvements and the average number of CV folds used.

Benchmark	#α	Median ΔFS	Mean ΔFS	Min ΔFS	Max ΔFS	#Folds
NQ-OPEN	5	0.206	0.192	0.151	0.253	40
NQ-OPEN-VEND	5	0.112	0.107	0.086	0.144	40
ASQA	5	0.092	0.089	0.071	0.109	40
AMBIGQA	5	0.075	0.074	0.056	0.094	40
HOTPOTQA	5	0.068	0.068	0.052	0.086	40
AMBIGQA-ENT	5	0.051	0.052	0.039	0.067	40

Experiment 3 Conformal Alignment Across benchmarks, the largest median and mean ΔFS arise on NQ-OPEN and NQ-OPEN-VEND, indicating that the gate is most effective precisely in the hardest, low-diversity factoid regimes. Moreover, all panels maintain positive min–max ranges over α, so factuality severity consistently drops on the kept set with no reversals. The proximity of median and mean within each benchmark suggests conformal alignment's stability across tolerance levels, i.e., the effect is not sensitive to α.