EMPIRICAL-BAYES TRUTH TILTING FOR CONFORMAL INFERENCE IN LARGE LANGUAGE MODEL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as answer engines, yet hallucinations make it important to quantify uncertainty in their answers. Open-ended question answering (QA) is challenging because repeated responses form prompt-specific semantic clusters and the true answer may remain outside the observed alphabet. Existing query-only conformal methods estimate missing sampling mass one prompt at a time and largely use generation frequency as a proxy for truth probability. We propose an empirical-Bayes framework that borrows strength across these local alphabets without aligning answer identities. A Pitman–Yor process (PYP) supplies a population-fitted cold-start reference for observed and unseen sampling mass; a cross-prompt truth-tilting layer uses labeled historical prompts to map sampling behavior and prompt features to posterior truth mass; and split conformal prediction provides finite-sample marginal validity. We expose reference–tilt non-identifiability and evaluate the components under matched supervision. Across eight short-answer settings, truth-supervised scores substantially improve negative log-likelihood and common-group reliability over query-only occupancy scores, whereas Good–Turing and PYP references are similar at budgets $T = 10, 20$. Separate short-prefix experiments show that integrating PYP residual-mass uncertainty improves future-count proper scores over an identical-mean plug-in forecast, with gains over flexible occupancy smoothing confined to scarce-history regimes. Truth tilting is therefore the main fixed-budget borrowing-strength mechanism, while PYP provides a principled cold-start component.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as answer generators in open-ended question answering (QA), yet hallucinations (Farquhar et al., 2024; Huang et al., 2025) and semantically diverse or contradictory responses under repeated sampling (Manakul et al., 2023; Kuhn et al., 2023) make reliable uncertainty quantification (UQ) essential. Among existing approaches, conformal prediction (CP; Vovk et al., 2005; Shafer & Vovk, 2008) is especially attractive because it constructs a prediction set with a rigorous distribution-free finite-sample coverage guarantee under exchangeability. Its practical value, however, depends on the nonconformity score that orders candidate answers and on whether the resulting set is informative for heterogeneous prompts.

Open-ended QA departs from standard classification because each prompt induces its own label space. Semantic clustering merges equivalent sampled responses into answer clusters, but cluster identities are not shared across prompts (Kuhn et al., 2023; Aichberger et al., 2025). Under a finite query budget T, the true answer may not belong to any observed cluster. Conformal Prediction with Query Oracle (CPQ; Noorani et al., 2025) represents this beyond-seen event by an "everything else" (EE) atom and writes the local alphabet as $\mathcal{Y}_T(X)$.

After augmenting the observed clusters with EE, a natural score uses estimated probability mass on the local alphabet. CPQ obtains these masses from the prompt-wise Good–Turing (GT) estimator (Good, 1953; Gale & Sampson, 1995). Two obstacles remain. First, purely local frequency-of-frequency estimates are unstable when T is small. Second, the probability that an LLM samples an answer is not the probability that the answer is correct. Ten responses to "What is the capital of Australia?" may

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

contain eight instances of `Sydney` and two of `Melbourne`, while the correct `Canberra` remains unseen. Better estimation of this sampling law only makes the wrong target more precise.

The prompt-by-prompt problem has a classical normal-means analogue. Coordinatewise maximum likelihood uses only each parameter's observation, whereas empirical Bayes and James–Stein shrinkage estimate population structure and let apparently unrelated coordinates inform one another (Stein, 1956; James et al., 1961; Efron, 2012). For at least three coordinates, James–Stein shrinkage can uniformly reduce total squared-error risk. Likewise, prompts have disjoint answer identities yet share task- and generator-level regularities in occupancy, error modes, and truth behavior.

We borrow strength at two levels. A Pitman–Yor process (PYP; Pitman, 1996; Pitman & Yor, 1997) learns species parameters from historical occupancy partitions and provides an analytic cold-start reference over observed clusters and residual mass. In the prior-washout intuition associated with Bernstein–von Mises asymptotics (van der Vaart, 1998), local evidence eventually dominates the population prior and Bayesian summaries approach frequentist estimators. This is intuition rather than a theorem for our clustered PYP model; empirically, PYP and Good–Turing references are already similar at $T = 10, 20$. Labeled historical prompts separately train the truth tilt, which transfers a sampling-to-truth correction and remains central even after the sampling law stabilizes.

The resulting empirical-Bayes procedure, BAYES-STATE, reweights a fixed local reference into a truth-aware score and calibrates it on a disjoint split; no reference answer is required for a test prompt. Because an unrestricted reference–tilt factorization is not identifiable, truth supervision determines the normalized score, while fixed-reference and matched-supervision comparisons identify the practical role of the reference. Split conformal calibration restores marginal validity; score quality governs conditional behavior, fallback use, and efficiency.

Figure 1 summarizes the workflow. Our contributions are:

1. We formulate empirical-Bayes learning over prompt-specific semantic alphabets and introduce truth tilting, which transfers sampling-to-truth structure without sharing answer identities.
2. We use a PYP species model as a principled cold-start reference with a residual-mass predictive law, and expose reference–tilt non-identifiability through matched supervision and channel ablations.
3. We prove marginal conformal validity and an excess mean-squared conditional-error bound, then evaluate posterior quality, absolute conformal performance, fallback use, and cold-start uncertainty across eight short-answer settings.

1.1 RELATED WORK

Conformal prediction and conditional validity. CP guarantees marginal coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018). Nontrivial exact distribution-free conditional validity is generally unavailable without additional assumptions (Vovk, 2012; Lei & Wasserman, 2014). Methods therefore target improved conditional behavior (Romano et al., 2019; Kiyani et al., 2024; Plassier et al., 2025; Chen & Li, 2025) or relaxed group/local notions (Jung et al., 2023; Guan, 2023; Hore & Barber, 2025; Gibbs et al., 2025). Density- and distribution-based scores connect estimated truth probabilities to set construction (Chernozhukov et al., 2021; Izbicki et al., 2022). We use conditional diagnostics alongside, not in place of, coverage and efficiency: lower groupwise conditional-coverage error need not mean smaller or more useful sets.

Conformal prediction for LLMs. Relevant work covers QA (Quach et al., 2024; Kladny et al., 2025; Noorani et al., 2025; Wang et al., 2026), factuality (Mohri & Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025; Jiang et al., 2025; Hu et al., 2026), and risk-controlled generation (Chen et al., 2025; Overman & Bayati, 2025; Valkanas et al., 2025). Conformal Language Modeling (CLM) calibrates sampling and rejection using an admission function (Quach et al., 2024); Sequential Conformal Prediction for Generative Models (SCOPE-Gen) uses greedy filtering (Kladny et al., 2025); CPQ connects query allocation and beyond-seen prediction to missing mass (Noorani et al., 2025); Risk-Constrained Sample-then-Filter (SAFER) separates sampling and filtering failure (Wang et al., 2026). We retain sampled candidates and `EE`, while learning a shared truth correction for the score. The contribution is a shared truth-score construction on local alphabets and its component-level evaluation, not the sampling–truth distinction or admission calibration alone.

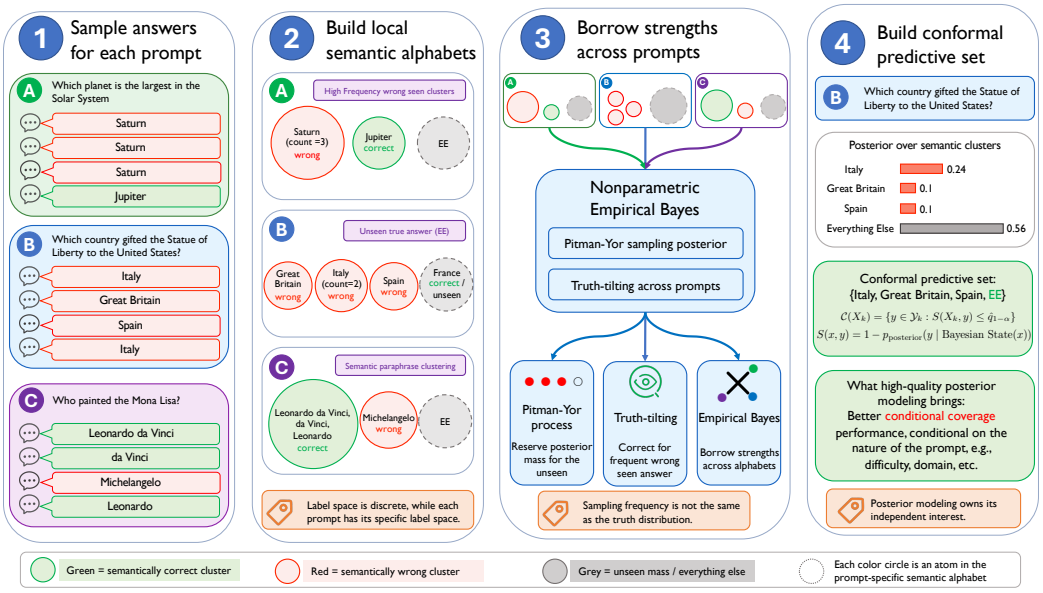

Figure 1: Overview of BAYES-STATE. A Pitman–Yor posterior supplies cold-start sampling mass, empirical-Bayes truth tilting corrects sampling behavior using labeled prompts, and split conformal prediction calibrates the resulting local score.

2 BASIC FORMULATION

Local semantic alphabets. Let X be a prompt and $Y^\star$ its ground-truth answer. Given a fixed query budget T, a black-box LLM returns strings $Z_{1:T}(X)$. A prompt-dependent semantic map κ_X clusters these strings into observed atoms $\mathcal{A}_T(X) = \{a_1, \dots, a_K\}$ with counts $n_k = \#\{j : \kappa_X(Z_j) = a_k\}$. The clustering map can be implemented by normalized exact matching, embedding clustering, bidirectional entailment, or an LLM-based semantic judge, depending on the QA task.

Finite querying may miss the true answer, so we augment the observed alphabet by an everything-else atom and write $\mathcal{Y}_T(X) = \mathcal{A}_T(X) \cup \{\texttt{EE}\}$. The collapsed truth atom of $Y^\star$ is

$$A_T^\star(X, Y^\star) = \begin{cases} a_k, & \kappa_X(Y^\star) = a_k \text{ for some } k, \\ \texttt{EE}, & \text{otherwise.} \end{cases} \tag{1}$$

A prediction set is a subset $C_T(X) \subseteq \mathcal{Y}_T(X)$, and atomic coverage is $A_T^\star \in C_T(X)$. In open-ended QA, the admission rule may deem multiple sampled atoms acceptable; we write $G_T(X, Y^\star) \subseteq \mathcal{Y}_T(X)$ and use the coverage event $C_T(X) \cap G_T(X, Y^\star) \neq \emptyset$. The singleton case has $G_T = \{A_T^\star\}$. If no observed atom is acceptable, set $G_T = \{\texttt{EE}\}$; otherwise $G_T \subseteq \mathcal{A}_T(X)$.

For a designated single reference target, the truth law is categorical on the local alphabet. Rather than conditioning only on the raw text X, we use the queried QA state $B_T = B_T(X)$, which summarizes the observed atoms, their counts, and global sampling behavior; see Appendix A.1 for details. We estimate

$$\eta_T^\star(a \mid B_T) = \mathbb{P}(A_T^\star = a \mid B_T), \qquad a \in \mathcal{Y}_T(X).$$

This posterior is distinct from the LLM sampling law: sampling mass describes where the model tends to generate, while truth mass describes where the reference answer lies. We approximate this law with a normalized truth score. For multi-atom admission, the score need not be a joint posterior over acceptable sets; the marginal intersection guarantee remains applicable.

Pitman–Yor process. The Pitman–Yor process is a prior over discrete sampling laws with an unbounded number of species. For prompt x, let G_x be the LLM law over semantic atoms: $G_x(\mathcal{U})$ is the probability that another response falls in $\mathcal{U}$. It describes generation, not correctness.

We use $G_x \sim \mathrm{PY}(\sigma, \theta, H)$ with diffuse base H, discount $0 \le \sigma < 1$, and strength $\theta > -\sigma$ (with $\theta > 0$ when $\sigma = 0$). Neither the number nor the identities of species are fixed a priori; Appendix C.1 derives the partition and residual-mass laws.

Given the first T clustered samples, suppose we observe semantic atoms $\mathcal{A}_T(x) = \{a_1, \ldots, a_K\}$ with counts $n_1, \ldots, n_K$. Integrating out G_x gives the Pitman–Yor posterior-predictive masses

$$\omega_{\psi,T}(a_k \mid B_T) = \frac{n_k - \sigma}{T + \theta}, \qquad \omega_{\psi,T}(\mathrm{EE} \mid B_T) = \frac{\theta + \sigma K}{T + \theta}, \qquad \psi = (\sigma, \theta). \tag{2}$$

The second term aggregates all unseen species without enumerating their identities. It provides a population-fitted cold-start reference when local occupancy data are sparse; as T grows, prompt-specific counts dominate, while correctness still depends on the separately learned truth tilt.

Coverage objectives. Let $\tau = 1 - \alpha$ and let $H_T = \mathbb{1}\{A_T^\star \in C_T(X)\}$ be the atomic coverage indicator; in the set-valued case, replace it by $H_T = \mathbb{1}\{C_T(X) \cap G_T(X, Y^\star) \neq \emptyset\}$. Split conformal calibration targets marginal coverage, $\mathbb{E}[H_T] \ge \tau$, but this population guarantee can hide substantial heterogeneity: easy regions may be over-covered while difficult regions remain under-covered.

We therefore evaluate conditional coverage together with marginal coverage, candidate count, and fallback use. The prompt-conditional coverage function and its mean squared conditional error (MSCE; Kiyani et al., 2024; Chen & Li, 2025) are

$$\pi_X(x) = \mathbb{P}(H_T = 1 \mid X = x), \qquad \mathrm{MSCE}_{\mathrm{X}} = \mathbb{E}[(\pi_X(X) - \tau)^2]. \tag{3}$$

The identity $\mathbb{E}[\pi_X(X)] = \mathbb{E}[H_T]$ separates marginal coverage from conditional heterogeneity. We evaluate both, together with set efficiency.

Bayesian state. The full queried state B_T contains X and its sampling behavior; the fitted model uses the compact features in Appendix A.1.

Let $\pi_B(B_T) = \mathbb{P}(H_T = 1 \mid B_T)$ and $\mathrm{MSCE}_{\mathrm{B}} = \mathbb{E}[(\pi_B(B_T) - \tau)^2]$. Since the full queried state refines the raw prompt, i.e., $\sigma(X) \subseteq \sigma(B_T)$, Proposition D.4 gives the orthogonal decomposition

$$\mathrm{MSCE}_{\mathrm{B}} = \mathrm{MSCE}_{\mathrm{X}} + \mathbb{E}[(\pi_B(B_T) - \pi_X(X))^2], \tag{4}$$

Hence $\mathrm{MSCE}_{\mathrm{X}} \le \mathrm{MSCE}_{\mathrm{B}}$. The same contraction applies to groups measurable from B_T. This requires the full state to contain X: coverage conditional on compressed features alone does not control arbitrary prompt-conditional coverage.

Remark 2.1 (Why a sampling-behavior state). The formal B_T contains X, whereas fitting uses low-dimensional occupancy, entropy, missing-mass, atom-mark, and optional embedding features. Similar states may share evidence about truth absence or frequent wrong answers despite unrelated strings, but this borrowing assumes stability within a task/LLM population; compressed features need not be sufficient for arbitrary prompt-conditional coverage.

3 METHODOLOGY

We first learn a truth-aware score on each local alphabet, then calibrate it on disjoint labeled prompts. The reference model and the truth correction serve different statistical targets.

3.1 SHARED TRUTH TILTING WITH A FIXED REFERENCE

Let $b_T(a \mid B_T) > 0$ be a fixed reference on $\mathcal{Y}_T(X)$. It may be a PYP predictive, a stabilized Good–Turing law, or a uniform reference. A shared atom-type-aware function acts on count/rank, atom marks, occupancy summaries, and optional prompt features $r_T(a, B_T)$:

$$\eta_{b,\phi,T}(a \mid B_T) = \frac{b_T(a \mid B_T) \exp\{g_\phi(r_T(a, B_T))\}}{\sum_{a' \in \mathcal{Y}_T(X)} b_T(a' \mid B_T) \exp\{g_\phi(r_T(a', B_T))\}}. \tag{5}$$

The function is shared, but normalization is local. Permuting observed atoms and their features permutes the probabilities. Historical labels therefore transfer a state-to-truth relationship, not answer identities. With $b_T = \omega_{\widehat{\psi},T}$, Eq. (5) recovers BAYES-STATE. Zero reference masses require a declared positive stabilization; otherwise no finite tilt can recover their support.

Identification and a reference convention. Sampling observations alone do not identify correctness. Even truth supervision cannot uniquely separate an unrestricted reference from its correction.

Proposition 3.1 (Reference–tilt equivalence)**.** *For positive references b, b' and an unrestricted tilt, $g' = g + \log b - \log b' + c(B)$ gives the same normalized score as g. With fixed b and a strictly positive categorical truth law $\eta^\star$, imposing $g(\texttt{EE}, B) = 0$ gives, for each atom a,*

$$g(a, B) = \log \frac{\eta^\star(a \mid B)}{b(a \mid B)} - \log \frac{\eta^\star(\texttt{EE} \mid B)}{b(\texttt{EE} \mid B)}.$$

The proof and zero-support limits are in Appendix D.1. Fixing b before truth fitting removes the arbitrary decomposition from the target; it does not establish finite-sample identifiability of neural parameters or eliminate misspecification. Subtracting $g(\texttt{EE}, B)$ changes no predictions, so this convention does not require rerunning existing scores. A finite feature class need not contain every transformed g'; matched reference swaps test that practical dependence.

Fitting and what the experiments estimate. For an interpretable PYP reference, estimate ψ from historical occupancy partitions without correctness labels, then freeze it. This is the cold-start empirical-Bayes step. The truth correction uses a separate role for labels: across fitting prompts $i \in \mathcal{I}_{\text{fit}}$, the ideal categorical target is cross entropy, extended to the admission surrogate

$$\widehat{\phi} \in \arg\min_{\phi} - \sum_{i \in \mathcal{I}_{\text{fit}}} \log \sum_{a \in G_i} \eta_{b,\phi,T_i}(a \mid B_{T_i}). \tag{6}$$

For singleton G_i, this is proper categorical log loss. With several acceptable atoms, it is neither a full likelihood of G_i nor sufficient to identify how truth mass is divided within it. We claim an admission score in that case, not a uniquely recovered categorical truth posterior.

The fixed-budget experiments use the existing preliminary score fit and two balanced classification heads, rather than exact minimization of Eq. (6); Appendix A.2 specifies the objectives and prediction rule. Their supervised preliminary parameters are not truth-free species estimates. The separate cold-start study fits the sampling law from occupancy data alone. All fitting precedes calibration.

Although PYP is Bayesian nonparametric and Eq. (5) may use a nonparametric function class, the evaluated logistic and finite-dimensional multilayer perceptron (MLP) heads are parametric. Shared truth supervision supplies the central borrowing mechanism. PYP, GT, or a uniform law can serve as the fixed reference, so empirical Bayes describes cross-prompt estimation rather than a claim that only one reference or fitting rule can borrow strength.

3.2 High-probability semantic score and conformalization

Given $\widehat{\eta}_T$, define the missing-aware high-probability score

$$\widehat{S}_T(a_k) = 1 - \widehat{\eta}_T(a_k \mid B_T), \qquad \widehat{S}_T(\texttt{EE}) = 2 - \widehat{\eta}_T(\texttt{EE} \mid B_T). \tag{7}$$

The observed-atom score is standard in classification (Angelopoulos et al., 2024); the shift for EE follows CPQ (Noorani et al., 2025) and defers fallback relative to seen answers. It constrains the set geometry rather than guaranteeing efficiency for an arbitrary fitted posterior.

For conformal calibration, abbreviate

$$\mathcal{Y}_i = \mathcal{Y}_T(X_i), \qquad A_i^\star = A_T^\star(X_i, Y_i^\star), \qquad \widehat{S}_i(a) = \widehat{S}_T(a).$$

After empirical-Bayes fitting, the score $\widehat{S}_i$ is frozen. The calibration score is

$$R_i = \widehat{S}_i(A_i^\star), \qquad i \in \mathcal{I}_{\text{cal}}.$$

With $n_{\text{cal}} = |\mathcal{I}_{\text{cal}}|$, set $k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$ and $\widehat{q} = R_{(k)}$, where $R_{(k)}$ is the k-th order statistic of the calibration scores with the convention $R_{(n_{\text{cal}}+1)} = \infty$. For a new prompt, the prediction set is

$$\widehat{C}_T(X_{n+1}) = \{a \in \mathcal{Y}_{T_{n+1}}(X_{n+1}) : \widehat{S}_{n+1}(a) \le \widehat{q}\}.$$

The fitted model supplies a truth-aware score; the split-conformal step supplies the distribution-free marginal guarantee. For set-valued admission, use $R_i = \min_{a \in G_i} \widehat{S}_i(a)$; the intersection guarantee is proved in Appendix D.3.

Theorem 3.2 (Marginal validity)**.** *Assume that, conditional on the fitting split, the calibration examples and the test example are exchangeable after including the LLM query randomness, semantic clustering, and collapsed truth atom. Then*

$$\mathbb{P}\Big(A^\star_{n+1} \in \widehat{C}_T(X_{n+1})\Big) \geq 1 - \alpha.$$

Equivalently, the prediction set covers the true semantic atom, where $A^\star_{n+1} = \texttt{EE}$ when the truth is not represented by any observed semantic cluster.

The probability averages over calibration data and the new queried instance; it is not a per-prompt guarantee. The result holds for any fixed reference and misspecified tilt. Score quality therefore affects conditional behavior and efficiency, not the validity of the rank argument.

3.3 Posterior error and conditional-coverage limits

We now connect truth-score approximation to conditional coverage in the single-target formulation. Let $B = B_T(X)$ denote the Bayesian state of a queried QA instance. Since our posterior is built on B rather than directly on the raw prompt X, we state the result at the state-conditional level; as discussed in Section 2, controlling state-conditional MSCE also controls prompt-conditional MSCE by coarsening. The state B induces the local alphabet $\mathcal{Y}_B = \mathcal{A}(B) \cup \{\texttt{EE}\}$.

Let $A^\star \in \mathcal{Y}_B$ be the collapsed truth atom, with true state-conditional law $\eta^\star_B = \mathcal{L}(A^\star \mid B)$. Denote its fitted approximation by $\widehat{\eta}_B$. For any posterior η_B on $\mathcal{Y}_B$, let S_η be the high-probability score in Eq. (7), and define $C_{\eta,q}(B) = \{a \in \mathcal{Y}_B : S_\eta(B,a) \leq q\}$. Let $\tau = 1 - \alpha$, and let $\mathcal{K}(\widehat{\eta})$ denote expected Kullback–Leibler (KL) divergence. Write

$$\mathrm{MSCE}_{\mathrm{B}}(\eta, q) = \mathbb{E}\Big[\big(\eta^\star_B(C_{\eta,q}) - \tau\big)^2\Big], \qquad \mathcal{K}(\widehat{\eta}) = \mathbb{E}[D_{\mathrm{KL}}(\eta^\star_B \,\|\, \widehat{\eta}_B)].$$

For threshold q and tolerance $\xi \geq 0$, define the oracle boundary mass

$$\Gamma_B(q,\xi) = \eta^\star_B(\{a \in \mathcal{Y}_B : |S_{\eta^\star}(B,a) - q| \leq \xi\}).$$

Theorem 3.3 (Posterior accuracy controls excess MSCE)**.** *Condition on the fitted score and threshold q, with a fresh test observation retaining the stated conditional truth law. Suppose that there exists $L < \infty$ such that $\Gamma_B(q,\xi) \leq L\xi$ for all relevant B and ξ. Then*

$$\sqrt{\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q)} \leq \sqrt{\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q)} + \sqrt{2}\, L \sqrt{\mathcal{K}(\widehat{\eta})}.$$

We prove Theorem 3.3 in Appendix D.4, and give a finite-sample MSCE certificate in Appendix D.5. The theorem compares the learned high-probability score with its oracle counterpart based on the true posterior. Posterior accuracy does not imply zero MSCE; instead, it controls the excess MSCE relative to the oracle score. The boundary-mass condition is a margin condition: little truth mass lies near the score threshold, so posterior estimation errors rarely change set membership. The oracle term is set-construction/discretization error, not irreducible estimation error: with oracle masses $(0.7, 0.3)$, a deterministic nested set cannot have mass 0.6. The bound is not a monotonicity theorem asserting that every reduction in average negative log-likelihood improves empirical group coverage.

Two truth-tilting channels. The score distinguishes truth absence from ranking within the observed support. If $\widehat{q} \geq 1$, all observed atoms pass Eq. (7); with positive jitter below 10^{-3}, $\widehat{q} \geq 1.001$ is sufficient. A common-rule channel audit over 240 units and both GT/PYP preliminary references found full and gate-only sets identical in all 480 primary-target comparisons. Across the primary and sensitivity targets, 832/960 comparisons remain identical; the 128 changes occur at the lower DeepSeek target. Thus the main operating points learn chiefly whether truth is absent, although seen ranking can affect decisions in other threshold regimes. Appendix B.3 gives the complete audit.

Remark 3.4. In this paper, we focus on the single-answer case, where each prompt has a collapsed truth atom $A^\star \in \mathcal{Y}_B$. For set-valued QA, the admission rule may return an acceptable set $G \subseteq \mathcal{Y}_B$. We then replace membership by intersection, $A^\star \in \widehat{C}(B)$ by $G \cap \widehat{C}(B) \neq \emptyset$, and use $\min_{a \in G} \widehat{S}(B,a)$ as the calibration score. The conformal rank argument extends directly; posterior-to-MSCE transfer requires the corresponding boundary and score-stability assumptions for the set-valued coverage functional; see Appendix D.3 and D.8.

The LLM fixes the sampling mechanism throughout this section. A change of generator or task population calls for renewed fitting and conformal calibration.

4 EXPERIMENTS

We retain the original fixed-budget experiments and separate three questions: whether shared labels improve truth estimation, whether the sampling reference adds value under matched supervision, and which score channel changes conformal decisions. The archived tables keep their original evaluation rules, while a common-rule matrix supplies component attribution.

4.1 BASIC SETTINGS

Data. We select PopQA (Mallen et al., 2023) and Natural Questions Open (NQ-Open; Lee et al., 2019) as our evaluation benchmarks. Both are standard open-domain short-answer QA datasets, covering factual questions with compact answers. For each benchmark, we randomly sample 1,000 prompts for efficient evaluation. We use cached responses from DeepSeek-V4-Flash (DeepSeek-AI, 2026) and Mixtral-8x7B-Instruct (Jiang et al., 2024) under fixed query budgets. Question embeddings from Qwen3-Embedding-8B (Zhang et al., 2025) support ablations and conditional diagnostics.

The main comparison uses 500 historical labeled prompts and 500 test prompts. CPQ assigns all 500 labels to calibration; BAYES-STATE uses 250 for fitting and 250 for calibration, matching the total label budget but not the labels' roles. Thirty repeated splits describe variability on reused problems rather than independent datasets. Budgets $T = 10, 20$ isolate score construction from query policy, with small-T results reported separately. The component audit reuses these cohorts and conditions intervals on saved fits, so it supports attribution rather than independent confirmation. A generator or task shift requires refitting and recalibration.

Evaluation metrics. Posterior diagnostics report floored negative log-likelihood (NLL) of acceptable-set mass, truth mass, Brier score, top-one accuracy (Acc@1), and mean reciprocal rank (MRR) on $\mathcal{Y}_T(X)$, with multi-atom definitions in Appendix A.8. Table headers use singleton notation. These metrics evaluate truth estimation rather than the next-generation sampling law.

Conformal tables retain the original *operational fallback* convention: selecting EE displays only fallback, interpreted as the full alphabet, and receives coverage credit; otherwise the displayed observed atoms must hit the admission set. *Displayed size* counts answers shown and is zero on fallback. It is not the cardinality of the literal threshold set $\widehat{C}_T$. Appendix A.8 distinguishes these events explicitly. We report operational coverage, EE frequency, displayed size, group-MSCE, and worst-slab coverage (WSC; Cauchois et al., 2021; Romano et al., 2020). Lower MSCE and higher WSC describe more balanced operational coverage, not automatically more informative answers.

The original partitions use behavior-plus-embedding (z) KMeans, behavior-only (r) KMeans, and diversity K/T bins. Some behavior features contain fitted η_{EE}, so z/r partitions and WSC can differ by method; they are descriptive diagnostics rather than fixed-subpopulation comparisons. The K/T grouping is method-independent at a common budget. This distinction and a frozen-group protocol for supplementary evaluation are explicit in Appendix A.8.

Semantic clustering. For each prompt, we normalize sampled answers and group exact matches into clusters. We then merge clusters whose representatives are judged semantically equivalent by DeepSeek-V4-Flash. Native prefix-specific assignments are available for $T = 10, 20$. These caches do not by themselves validate arbitrary shorter prefixes as online clusterings. The assignments are cached and shared across methods. Finally, we append EE to represent the event that the true answer is absent from all observed semantic clusters.

4.2 TRUTH ESTIMATION AND MATCHED ATTRIBUTION

The historical diagnostics compare empirical frequencies, Good–Turing, PYP occupancy, and the PYP+Tilt instantiation of BAYES-STATE. The retained DeepSeek tables remain in Appendix B.1; they show that PYP occupancy mainly protects a small tail from near-zero mass, whereas supervised tilting changes truth mass and ranking. Because those rows do not match truth supervision, they cannot attribute the effect to PYP.

Table 1 summarizes all eight dataset–generator–budget cells under a common 250/250/500 split and 30 repeats. GT and PYP occupancy use no fitting truth labels; GT+Tilt and PYP+Tilt share the

Table 1: Full matched-supervision attribution, equal-cell means over eight dataset–generator–budget settings. NLL/Brier are lower-is-better; Acc@1/MRR higher-is-better. Raw-MSCE uses common fit-only occupancy groups. "Disp." is the archived fallback-collapsed display count; literal seen size is identical across methods at the primary targets.

Method	NLL	Brier	Acc@1	MRR	Raw-MSCE	EE	Disp.
GT	2.696	.963	.418	.674	.0361	.168	1.262
PYP occupancy	2.000	.991	.390	.645	.0413	.163	1.246
GT+Tilt	1.515	.655	.599	.785	.0119	.183	1.640
BAYES-STATE (PYP+Tilt)	1.527	.658	.598	.785	.0119	.183	1.641
Direct-State	1.498	.651	.598	.782	.0125	.184	1.633

Table 2: Absolute retained results under the same total historical-label budget; entries are CPQ/BAYES-STATE. Op., EE, and Disp. denote operational coverage, fallback frequency, and displayed size. MSCE uses K/T bins; WSC is worst-slab coverage. Targets are .6 for DeepSeek and .5 for Mixtral.

Generator	Dataset	T	Op.	EE	Disp.	K/T-MSCE	WSC
DeepSeek	PopQA	10	.593/.597	.140/.154	1.466/1.783	.032/.015	.214/.263
		20	.595/.598	.110/.130	1.940/2.409	.028/.017	.229/.263
	NQ-Open	10	.595/.598	.132/.152	1.101/1.316	.013/.005	.309/.317
		20	.596/.599	.117/.130	1.223/1.573	.011/.005	.316/.329
Mixtral	PopQA	10	.502/.502	.182/.196	1.215/1.544	.044/.009	.091/.184
		20	.499/.502	.163/.182	1.483/1.956	.044/.012	.097/.184
	NQ-Open	10	.501/.498	.246/.269	.824/1.141	.056/.005	.184/.171
		20	.500/.495	.228/.245	.910/1.424	.046/.005	.197/.180

archived Protocol-L learner; Direct-State is a supervised structural control without a reference offset. The table reports equal-cell means; Appendix B.2 gives cellwise results and paired intervals.

Relative to PYP occupancy, PYP+Tilt lowers absolute NLL from 2.000 to 1.527, a paired reduction of 0.4722 (95% interval $[0.3965, 0.5455]$), showing the value of shared truth supervision. Under matched supervision, absolute NLL is 1.515 for GT+Tilt and 1.498 for Direct-State; PYP+Tilt is higher by 0.0120 $[0.0092, 0.0149]$ and 0.0298 $[0.0196, 0.0396]$, respectively, while ranking and common-group reliability remain close. Thus reference choice is secondary at $T = 10, 20$, consistent with cold-start shrinkage receding as local evidence accumulates. Direct-State is a structural control rather than an exact reference swap because it uses a different derived-feature schema.

4.3 EVALUATION ON CONFORMAL COVERAGE PERFORMANCE

We compare three methods: our method, BAYES-STATE; CPQ (Noorani et al., 2025), which uses the Good–Turing estimator under a fixed query budget; and a seen-only frequency-without-missing-mass baseline (FREQ-NO-MISSING), which uses empirical frequencies over observed clusters and ignores the beyond-seen event. For each setting, we choose a nominal coverage level above the marginal coverage achieved by FREQ-NO-MISSING, so that accounting for EE is necessary. CPQ is the closest matched score interface; CLM (Quach et al., 2024) and SCOPE-Gen (Kladny et al., 2025) also change sampling/filtering and have different admission-cost requirements. Our fixed-query comparison does not establish superiority to their complete procedures.

Table 2 reports absolute operational results for CPQ and BAYES-STATE; paired differences, full z/r diagnostics, alternative targets, and split standard deviations appear in Appendix B. Truth-aware learning lowers the method-independent K/T-group mean-squared conditional error in all eight retained settings, with higher fallback use and displayed size. Across the eight settings, EE frequency rises by 1.3–2.3 percentage points and displayed size by 0.215–0.514 answers; worst-slab coverage improves in six settings. These are tradeoffs rather than a Pareto improvement. At the primary targets, literal seen size equals K for every method, so display compression reflects the EE fallback convention rather than fewer literal candidates. Appendix B.3 reports both set semantics and the lower-target ranking sensitivity.

4.4 COLD-START SAMPLING SCOPE

The matched truth task above shows that PYP is not the main source of moderate-budget truth gains. We therefore evaluate it on the sampling-law quantity it is designed to regularize: residual-mass uncertainty after a short prefix. This truth-free study isolates its empirical-Bayes cold-start role without treating sampling uncertainty as correctness.

Cold start concerns the sampling law. A separate experiment distinguishes M historical occupancy-training questions from the online prefix budget t; it uses no correctness labels. Under the fitted PYP law, residual mass outside the current alphabet has posterior $U_t \sim \mathrm{Beta}(a, b)$, where $a = \theta + \sigma K_t$ and $b = t - \sigma K_t$. The next h responses outside that *fixed* alphabet follow a beta-binomial count law, not a distinct-species count. Relative to the same-mean binomial plug-in, integration inflates variance by $1 + (h-1)/(t+\theta+1)$, giving PYP a concrete short-prefix role even when one-step means are similar. Appendix C.1 gives the full derivation and proper-score identities.

Across the four core cells and $M = 5, 10, 25$, $t = 2, 3, 5$, $h = 5$, integration changes count NLL by -0.4174 (95% interval $[-0.4351, -0.3993]$) and ranked probability score by -0.0857 ($[-0.0891, -0.0824]$) relative to its identical-mean plug-in forecast, isolating uncertainty propagation rather than mean improvement. The direction holds in all six evaluated dataset–generator cells and both alphabet constructions. Against a fitted occupancy beta-binomial, the scarce-history $M = 5$ aggregate improves by 0.2079 NLL and 0.0158 RPS, but both scores reverse from $M = 25$; exact-alphabet RPS favors PYP in only $5/10$ overlapping training panels. One-step comparisons also fail the practical screen against stabilized GT and an occupancy cross-validated control, while PYP intervals are wider. The supported claim is therefore a coherent empirical-Bayes cold-start law with a bounded scarce-history advantage, not uniform estimator, truth, conformal, or policy superiority. Appendix C.3 reports all budgets, alphabets, intervals, and TriviaQA sensitivities.

4.5 SCOPE, SUPERVISION, AND DEPLOYMENT

The empirical claim is limited to short-answer QA within the sampled task/LLM populations and the cached admission rules. Each benchmark supplies 500 historical labeled prompts. CPQ spends all 500 on calibration, whereas the fixed-budget BAYES-STATE comparison uses 250 for tilt fitting and 250 for calibration. The common-rule matrix instead fixes the latter 250/250 allocation for every method, so it answers component attribution rather than label-budget allocation. At test time, the method receives only the prompt, sampled responses, their clusters, the frozen score model, and the calibrated threshold—never a reference answer. A changed generator, domain, or admission rule requires refitting and recalibration; long-form semantic clustering remains unvalidated.

Generation and clustering dominate cost and are shared with query-based baselines. Given clusters, PYP costs $O(K)$ and the tilt evaluates at most $T + 1$ atoms. An archived run took 26.36 seconds to fit (2.42 for the preliminary empirical-Bayes fit and 23.94 for the heads) and 0.104 seconds to calibrate, versus 0.028 seconds for CPQ. These timings are implementation-specific, and the retained fixed-budget comparisons use fallback more often and display larger sets.

5 CONCLUSION

We presented empirical-Bayes truth tilting for conformal QA on prompt-specific alphabets. The analysis separates three statistical tasks that query-only methods tend to conflate: estimating a sparse local sampling law, learning which sampling states indicate truth, and calibrating a valid prediction set. Historical labels identify the normalized truth-aware score rather than a unique reference–tilt decomposition; matched controls show that cross-prompt truth supervision, not the PYP offset alone, drives the main fixed-budget gains. PYP retains a principled role at cold start, where its residual-mass posterior propagates uncertainty missed by a plug-in mean, while its influence recedes as local evidence accumulates in the prior-washout regime. Split conformal calibration then guarantees marginal validity without requiring a reference answer for a new prompt. Conditional reliability, fallback cost, and transfer beyond the sampled task and generator populations remain empirical questions for future frozen-protocol studies.

Artificial intelligence use statement

Generative artificial intelligence (AI) tools assisted with proof checking, manuscript organization, language polishing, and formatting. Revision support also included experiment planning and code/result auditing. The authors are responsible for all claims, evidence, and final content.

Reproducibility statement

Appendix A specifies features, fitting, splits, and evaluation. Appendix C.2 gives the cold-start protocol; Appendix D contains the proofs. Cached-data reuse, model-version limits, and numerical conventions are explicit.

Ethics statement

The work evaluates uncertainty on public short-answer benchmarks. An `EE` response represents uncertainty or fallback, not a verified textual answer. Coverage does not establish factual accuracy for every question, and the method requires population-appropriate calibration before deployment.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Improving uncertainty estimation through semantically diverse language generation. In *International Conference on Learning Representations*, 2025.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Catherine Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. In *International Conference on Machine Learning*, 2025.

Qianyi Chen and Bo Li. Colorful pinball: Density-weighted quantile regression for conditional guarantee of conformal prediction. *arXiv preprint arXiv:2512.24139*, 2025.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, 2024.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report, Hugging Face, 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf`. Accessed: 2026-05-04.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

William A. Gale and Geoffrey Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87 (2):549–578, 2025.

Zirui Hu, Zheng Zhang, Yingjie Wang, Leszek Rutkowski, and Dacheng Tao. Cofact: Conformal factuality guarantees for language models under covariate shift. In *International Conference on Learning Representations*, 2026.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

William James, Charles Stein, et al. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 361–379. University of California Press, 1961.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Conformal linguistic calibration: Trading-off between factuality and specificity. In *Advances in Neural Information Processing Systems*, 2025.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *International Conference on Machine Learning*, 2024.

Klaus-Rudolf Kladny, Bernhard Schölkopf, and Michael Muehlebach. Conformal generative modeling with improved sample efficiency through sequential greedy filtering. In *International Conference on Learning Representations*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6086–6096, 2019.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, 2024.

Sima Noorani, Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction beyond the seen: A missing mass perspective for uncertainty quantification in generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *Advances in Neural Information Processing Systems*, 2025.

Jim Pitman. Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, 30:245–267, 1996.

Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *International Conference on Machine Learning*, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, and Surbhi Goel. Conformal language model reasoning with coherent factuality. In *International Conference on Learning Representations*, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*, pp. 197–207. University of California Press, 1956.

Antonios Valkanas, Soumyasundar Pal, Pavel Rumiantsev, Yingxue Zhang, and Mark Coates. C3po: Optimized large language model cascades with probabilistic cost constraints for reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

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

Qingni Wang, Yue Fan, and Xin Eric Wang. Safer: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Implementation Details

A.1 Bayesian State

For a prompt x, the first T sampled responses are clustered into observed semantic atoms $\mathcal{A}_T(x) = \{a_1, \ldots, a_K\}$ with counts $n_1, \ldots, n_K$. We write $\mathcal{Y}_T(x) = \mathcal{A}_T(x) \cup \{\texttt{EE}\}$, where EE is the unseen semantic remainder. Let $p_k = n_k/T$ and $N_r = \sum_{k=1}^{K} \mathbb{1}\{n_k = r\}$.

The queried state $B_T(x)$ contains the local alphabet, the occupancy counts, atom-level marks, and global sampling summaries. Its purpose is to summarize the finite-query sampling behavior through which the truth posterior is inferred.

The Pitman–Yor layer first defines a semantic sampling posterior

$$\omega_{\psi,T}(a_k \mid B_T) = \frac{n_k - \sigma}{T + \theta}, \qquad \omega_{\psi,T}(\texttt{EE} \mid B_T) = \frac{\theta + \sigma K}{T + \theta},$$

where $\psi = (\sigma, \theta)$ is learned by empirical Bayes. This distribution models where the LLM tends to sample next, not where the ground-truth answer lies. The empirical-Bayes truth posterior is then obtained by the general truth-tilting form used in the main text:

$$\eta_{\psi,\phi,T}(a \mid B_T) = \frac{\omega_{\psi,T}(a \mid B_T) \exp\{g_\phi(r_T(a, B_T))\}}{\sum_{a' \in \mathcal{Y}_T(x)} \omega_{\psi,T}(a' \mid B_T) \exp\{g_\phi(r_T(a', B_T))\}}, \qquad a \in \mathcal{Y}_T(x).$$

Here $r_T(a, B_T)$ is the feature vector exposed to the tilt, and g_ϕ may be a linear model, a logistic model, or a small multilayer perceptron (MLP). In the formulation, an atom-type-aware function can represent different corrections for seen atoms and EE. The archived two-head fitting procedure and its surrogate objectives are detailed in Appendix A.2.

We use a low-dimensional prompt-level behavior vector

$$u_T(B_T) = \big[T,\ K,\ K/T,\ p_{\max},\ \mathrm{Ent}(p_{1:K}),\ N_1/T,\ N_2/T,\ \eta_T^{\mathrm{base}}(\texttt{EE} \mid B_T)\big], \tag{8}$$

where $p_{\max} = \max_k p_k$ and $\mathrm{Ent}(p_{1:K}) = -\sum_{k=1}^{K} p_k \log p_k$ is the empirical entropy of the sampled semantic-cluster distribution, with the usual convention $0 \log 0 = 0$.

The entries of $u_T(B_T)$ have the following roles. T is the query budget; K is the number of observed semantic atoms; K/T measures semantic diversity; $p_{\max}$ measures concentration on the dominant cluster; $\mathrm{Ent}(p_{1:K})$ measures dispersion of sampled semantic mass; N_1/T and N_2/T are frequency-of-frequency summaries that are informative about the long tail and missing mass; and $\eta_T^{\mathrm{base}}(\texttt{EE} \mid B_T)$ is a preliminary truth-score feature, distinct from sampling missing mass and not presumed calibrated.

For each observed atom a_k, we use an atom-level feature vector

$$\phi_{T,k} = \big[\log(n_k + 1),\ p_k,\ \mathbb{1}\{n_k = 1\},\ \mathbb{1}\{n_k = 2\},\ \mathrm{rank}_k/K,\ \ell_k,\ \omega_{\psi,T}(a_k \mid B_T),\ \eta_T^{\mathrm{base}}(a_k \mid B_T)\big].$$

Here rank_k is the rank of a_k by decreasing count, ℓ_k is a normalized representative-answer length feature, $\omega_{\psi,T}(a_k \mid B_T)$ is the Pitman–Yor sampling mass, and $\eta_T^{\mathrm{base}}(a_k \mid B_T)$ is the preliminary local truth mass. The base posterior η_T^{base} is computed by a simple local tilt of $\omega_{\psi,T}$ using count and length features, and is used only as an auxiliary feature for the final truth-tilting model.

The final feature map is composed of a **atom-level** component plus a **prompt-level** component

$$r_T(a_k, B_T) = \big[\phi_{T,k};\ u_T(B_T);\ e(x);\ 0\big], \qquad r_T(\texttt{EE}, B_T) = \big[\phi_{T,\texttt{EE}};\ u_T(B_T);\ e(x);\ 1\big],$$

where $e(x) \in \mathbb{R}^{d_e}$ is an optional compressed prompt embedding ($d_e = 256$ in our experiments), and the last coordinate is an atom-type indicator. For EE, unavailable atom-specific entries in $\phi_{T,\texttt{EE}}$ are set to zero, while the available entries include $\omega_{\psi,T}(\texttt{EE} \mid B_T)$ and $\eta_T^{\mathrm{base}}(\texttt{EE} \mid B_T)$. The type indicator prevents EE from being treated as an ordinary zero-count seen atom.

The low-dimensional state focuses shared estimation on recurring occupancy and correctness patterns, rather than matching answer identities. Prompt embeddings are optional. Their added value depends on the fitted head and available labels; a compact representation is not assumed sufficient for the full prompt-conditional truth law.

Write $\widehat{\eta}_T = \eta_{\widehat{\psi},\widehat{\phi},T}$ after fitting. The main experiments use

$$\widehat{S}_T(a_k) = 1 - \widehat{\eta}_T(a_k \mid B_T), \qquad \widehat{S}_T(\texttt{EE}) = 2 - \widehat{\eta}_T(\texttt{EE} \mid B_T).$$

The shift for EE follows the CPQ-style beyond-seen construction: EE is a fallback atom for the unseen semantic remainder, not an ordinary observed answer cluster.

A.2 Implemented fitting and its statistical interpretation

The original tables use the archived implementation, not a newly fitted model. Its no-tilt Pitman–Yor empirical-Bayes (PYP-EB) arm uses an occupancy-partition likelihood to fit species parameters. The full score first uses a small supervised grid over reference and local correction parameters, then fits two shared classification heads on historical prompts. If several atoms are admitted, the preliminary grid scores the admitted atom with largest current mass; this differs from maximizing the sum of admitted masses. Consequently, those preliminary parameters are score-fitting quantities and should not be interpreted as separately identified sampling-law parameters.

Writing $q_\phi(B)$ for the fitted absence classifier and $h_\phi(r_k)$ for the seen classifier log-odds, the implemented two-head prediction has the form

$$\widehat{\eta}(\texttt{EE} \mid B) = q_\phi(B), \qquad \widehat{\eta}(a_k \mid B) = (1 - q_\phi(B)) \frac{\eta^{\text{base}}(a_k \mid B) e^{h_\phi(r_k)}}{\sum_j \eta^{\text{base}}(a_j \mid B) e^{h_\phi(r_j)}}.$$

The configurable implementation can blend the gate with the preliminary score; the state-only control removes reference-derived features and uses no reference offset. This decomposition identifies the two practical truth-learning channels while remaining a normalized local score.

The logistic implementation standardizes fitting features and uses two binary classifiers with $C = 1$, at most 1,000 iterations, and balanced class weights. The optional MLP has hidden widths $(32, 16)$; both are finite-dimensional learners. Class weighting changes the binary fitting population, and odds of a correctness classifier are not automatically the truth-to-sampling density ratio. These are practical surrogate fits, not exact optimization of the normalized admission objective in Section 3.1. Marginal conformal validity requires a frozen score and exchangeability, not correctness of this posterior approximation. A likelihood-corrected implementation would be a separate estimator and requires its own experimental row.

The linear seen head concatenates local and global covariates. A global additive term common to all atoms cancels under seen softmax; context-dependent changes of their relative ordering require interactions or a nonlinear head. Prompt context can still affect the absence head. Atom ranks/ties must carry the same marks under a relabeling; the archived first-arrival tie convention should not be mistaken for invariance to changing the observed response sequence.

For sampling attribution, fit PYP on truth-free partitions and freeze it before truth correction. This avoids identifying both factors through one flexible truth loss. The archived full-score numbers retain their original fitting protocol.

A.3 Semantic clustering and truth labels

For each prompt, we convert sampled responses into a prompt-specific semantic alphabet before fitting posteriors or applying conformal calibration. We first apply task-specific answer normalization and exact-match grouping. For arithmetic tasks, we extract the final numeric answer. For short-answer QA, including PopQA, TriviaQA, and NQ-Open, we remove common answer wrappers such as "the answer is", punctuation, articles, and extra whitespace, and then normalize case. The exact-match clusters form an initial partition of the sampled responses.

For the main semantic-clustering experiments, we refine these exact-match clusters using a cached LLM equivalence judge. For each prompt and prefix $T \in \{10, 20\}$, we compare representatives of the exact-match clusters and merge clusters judged semantically equivalent. The judge is DeepSeek-V4-Flash with deterministic decoding. We use the following pairwise prompt:

```
Question:  <question>
Answer A: <cluster representative A>
Answer B: <cluster representative B>
Do Answer A and Answer B express the same answer to the question?
Reply with exactly YES or NO.
```

Only the representatives of exact-match clusters are compared, rather than all raw responses. Within each materialized prefix, later clusters are greedily merged into earlier semantically equivalent clusters. Native $T = 10, 20$ assignments are available and their restrictions have been checked for consistency. That check alone is not a provenance proof for arbitrary online prefixes; shorter retrospective slices are labeled as such. All pairwise judgments are stored in a local pair cache, and the resulting cluster assignments are written into the response cache under prefix-specific keys. All methods in a given experiment use the same cached semantic alphabet. A cluster is marked acceptable if it contains a cached exact-match anchor to the reference/aliases. This operational rule may mix correct and incorrect responses in a semantic cluster and is not independent human semantic gold. Multiple acceptable clusters are retained. The special atom `EE` is appended after clustering and denotes that no observed semantic cluster is acceptable under the task admission rule.

A.4 Response generation and prompt embeddings

We use answer-only sampling prompts to reduce formatting noise in exact matching and semantic clustering. The system message is:

```
You are a precise QA system.  Answer with only the final short answer
entity or phrase.  Do not include explanations.
```

Mixtral experiments use `mistralai/mixtral-8x7b-instruct`; DeepSeek experiments use the official DeepSeek application programming interface (API) with `deepseek-v4-flash`. Unless otherwise specified, responses are sampled with temperature 0.7, top-$p = 0.95$, and a short-answer maximum token budget. Response generation is cached with a sidecar file so interrupted runs can resume without resampling completed responses. Prompt embeddings are generated once using `qwen/qwen3-embedding-8b` at 256 dimensions and reused across LLMs, since they depend only on the question. Historical API aliases and embedding identifiers lack immutable revision/timestamp evidence in the caches; new generations must form a separate cohort rather than be appended as if produced by the same frozen model.

For PopQA and NQ-Open, we use 1,000 validation prompts sampled by a fixed-seed streaming shuffle. For each prompt, we generate 20 responses in the main DeepSeek experiments and use fixed prefixes $T = 10$ and $T = 20$. For Mixtral TriviaQA, when a 50-response cache is available, the semantic clustering and experiments still materialize only the $T = 10$ and $T = 20$ prefixes, so later responses do not affect the smaller-prefix experiments.

A.5 Posterior estimators

The posterior-quality experiments compare four estimators over the local alphabet $\mathcal{Y}_T(x) = \{a_1, \ldots, a_K, \texttt{EE}\}$. Freq-no-missing uses empirical frequencies on observed clusters, $\hat{p}(a_k) = n_k/T$, and assigns zero mass to `EE`. Good–Turing is the CPQ-style frequency-of-frequencies estimator that reserves missing mass for `EE`. PYP-EB uses a Pitman–Yor posterior predictive distribution with empirical-Bayes fitted species parameters, but does not use truth tilting. Bayes-State is the full estimator: it starts from the Pitman–Yor predictive distribution and applies shared truth tilting using cluster-local features, sampling-behavior statistics, and prompt embeddings.

For the conformal experiments, CPQ uses the Good–Turing distribution with the CPQ-style score $S(a_k) = 1 - \hat{p}(a_k)$ and $S(\texttt{EE}) = 2 - \hat{p}(\texttt{EE})$. Bayes-State uses the same score form, replacing $\hat{p}$ with our fitted truth posterior $\hat{\eta}_T$. Freq-no-missing is included only as a seen-only diagnostic baseline: it never includes `EE`, and its coverage estimates the maximum attainable coverage using observed clusters alone.

A.6 CPQ-style score and tie handling

Because the `EE` score is highly discrete, especially when $\hat{p}(\texttt{EE})$ is near zero, we follow the CPQ implementation and add a small tie-breaking jitter to CPQ-style scores during calibration and prediction. Concretely, for observed clusters we use $S(a_k) = 1 - \hat{p}(a_k) + \epsilon_k$, and for `EE` we use $S(\texttt{EE}) = 2 - \hat{p}(\texttt{EE}) + \epsilon_{\texttt{EE}}$, with $\epsilon \in [0, 10^{-3}]$. In our implementation this jitter is generated deterministically from example and atom identifiers so that results are reproducible. The mathematical threshold set retains every passing atom. The archived display policy instead returns only `EE` whenever `EE` passes, and otherwise returns the passing observed atoms. For the operational evaluation, this

display is an abstention shorthand for the entire local alphabet, not a literal unseen-only singleton. Formally, let

$$\overline{C}_T = \begin{cases} \mathcal{Y}_T(X), & \texttt{EE} \in \widehat{C}_T, \\ \widehat{C}_T, & \texttt{EE} \notin \widehat{C}_T. \end{cases}$$

Then $\overline{C}_T \supseteq \widehat{C}_T$, so expansion preserves the marginal lower bound. Its conditional coverage and cardinality need not equal those of $\widehat{C}_T$. Selecting EE is not a correct textual answer. Tables retain operational coverage and display cost, not literal-set metrics.

A.7 Data splitting

Our final experiments use a two-stage split that avoids giving training-free baselines an unused fitting split. For FREQ-NO-MISSING and CPQ, the data are split into 50% calibration and 50% test. For methods that fit a Bayesian posterior, the same non-test half is split into 25% fitting and 25% conformal calibration, with the remaining 50% used for test. Thus CPQ is allowed to use all non-test labeled data for calibration, while Bayes-State uses half of the non-test data to fit the posterior and half to calibrate the score. Unless otherwise stated, we average results over repeated random splits with a fixed list of seeds.

A.8 Metrics

Posterior-quality metrics. For each test prompt i, let $\mathcal{Y}_i = \mathcal{A}_T(X_i) \cup \{\texttt{EE}\}$ be its local semantic alphabet and let $G_i \subseteq \mathcal{Y}_i$ be the set of acceptable truth atoms. In the single-answer case, G_i is a singleton; if no observed cluster is correct, then $G_i = \{\texttt{EE}\}$. Given an estimated posterior $\hat{p}_i(\cdot)$ over $\mathcal{Y}_i$, define the true-set mass $q_i = \hat{p}_i(G_i) = \sum_{a \in G_i} \hat{p}_i(a)$. We report

$$\mathrm{NLL} = \frac{1}{n}\sum_{i=1}^{n} -\log(q_i \vee \varepsilon), \qquad \widehat{p}(A^\star) = \frac{1}{n}\sum_{i=1}^{n} q_i,$$

where $\varepsilon = 10^{-12}$ is the archived numerical floor. A finite floored NLL for a zero-support estimator is not a finite untruncated log loss. For multi-atom G_i, $-\log \hat{p}_i(G_i)$ evaluates admission mass and is not the proper joint log loss of a random correct-support set. The Brier score is computed against the normalized set-valued target $y_i(a) = |G_i|^{-1}\mathbf{1}\{a \in G_i\}$:

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a \in \mathcal{Y}_i}\big(\hat{p}_i(a) - y_i(a)\big)^2.$$

Top-1 accuracy is

$$\mathrm{Acc@1} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\arg\max_{a \in \mathcal{Y}_i} \hat{p}_i(a) \in G_i\right\}.$$

For ranking quality, let $r_i = \min_{a \in G_i} \mathrm{rank}_i(a)$, where atoms are ranked in decreasing posterior mass and rank one is best. We report

$$\mathrm{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i}.$$

For the unseen event, let $Y_i^{\texttt{EE}} = \mathbf{1}\{G_i = \{\texttt{EE}\}\}$ and $m_i = \hat{p}_i(\texttt{EE})$. We record the average predicted unseen mass $\frac{1}{n}\sum_i m_i$, the empirical unseen rate $\frac{1}{n}\sum_i Y_i^{\texttt{EE}}$, and binary calibration metrics

$$\text{EE-NLL} = \frac{1}{n}\sum_{i=1}^{n} -\left[Y_i^{\texttt{EE}}\log(m_i \vee \varepsilon) + (1 - Y_i^{\texttt{EE}})\log((1 - m_i) \vee \varepsilon)\right],$$

$$\text{EE-Brier} = \frac{1}{n}\sum_{i=1}^{n}(m_i - Y_i^{\texttt{EE}})^2.$$

We also compute everything-else expected calibration error (EE-ECE) by binning m_i into B equal-frequency bins:

$$\text{EE-ECE} = \sum_{b=1}^{B}\frac{|I_b|}{n}\left|\frac{1}{|I_b|}\sum_{i \in I_b} m_i - \frac{1}{|I_b|}\sum_{i \in I_b} Y_i^{\texttt{EE}}\right|.$$

The EE-specific metrics are diagnostics for the binary unseen event; the main posterior-quality metrics are NLL, true-set mass, Brier score, Acc@1, and MRR over the full local alphabet.

Conformal prediction metrics. For a conformal prediction set $C_i \subseteq \mathcal{Y}_i$, define the coverage indicator

$$H_i = \mathbf{1}\{C_i \cap G_i \neq \emptyset\}.$$

This is *literal* coverage of the threshold set. In the archived tables the evaluated event instead is $H_i^{\mathrm{op}} = \mathbf{1}\{\texttt{EE} \in \widehat{C}_i \text{ or a passing seen atom is admitted}\}$, equivalently coverage of $\overline{C}_i$ from the preceding subsection. Thus the displayed "Op. Cov." is $n^{-1}\sum_i H_i^{\mathrm{op}}$. Let $D_i = \{\texttt{EE}\}$ on fallback and $D_i = \widehat{C}_i$ otherwise. The recorded display metrics are

$$\mathrm{EEFrac} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\texttt{EE} \in D_i\}, \qquad \mathrm{Size}_{\mathrm{disp}} = \frac{1}{n}\sum_{i=1}^{n} |D_i \setminus \{\texttt{EE}\}|.$$

The resulting displayed size is zero on abstention, while the expanded set contains all observed atoms. Literal size $|\widehat{C}_i \setminus \{\texttt{EE}\}|$ must be recomputed from saved thresholds and cannot be inferred from this display size alone. All retained main and appendix tables use the original operational convention.

Group-based MSCE. For the archived tables use H_i^{op} in the following formulas; the literal event is an alternative that requires separate evaluation. Let τ be the nominal coverage level. Given a partition $\mathcal{G}$ of the test examples, let I_g be the examples in group g, $n_g = |I_g|$, and

$$\widehat{\mathrm{Cov}}_g = \frac{1}{n_g}\sum_{i \in I_g} H_i.$$

The empirical group-MSCE is

$$\widehat{\mathrm{MSCE}}(\mathcal{G}) = \sum_{g \in \mathcal{G}} \frac{n_g}{n}\left(\widehat{\mathrm{Cov}}_g - \tau\right)^2.$$

We use three main partitions: z-KMeans, behavior-KMeans, and response-diversity bins. For z-KMeans, we cluster $z_i = [u(B_T(X_i)); e(X_i)]$, i.e., sampling-behavior statistics concatenated with the prompt embedding. For behavior-KMeans, we cluster only $u(B_T(X_i))$. For response-diversity MSCE, we bin K_i/T, where K_i is the number of observed semantic clusters. The number of KMeans groups and one-dimensional bins is controlled by `-msce-kmeans-k` and `-eval-group-bins`, respectively. If KMeans returns fewer distinct groups because of duplicated behavior states, we use the effective number of nonempty groups. The archived evaluator fits these groups on evaluation features and includes the method-specific $\eta_{\texttt{EE}}$; hence z/r groups can change across methods even without using truth labels in clustering. For a fixed-population comparison, use method-independent features and fit/freeze the groups on the fitting split. We do not retroactively reinterpret the old group values as results of that corrected protocol.

Worst-slice coverage. We approximate worst-slice coverage by random-projection slabs in a covariate space z_i, by default $z_i = [u(B_T(X_i)); e(X_i)]$. For each of M random unit directions v_m, we sort the projections $\langle v_m, z_i\rangle$ and scan contiguous slabs containing at least $\lceil \beta n \rceil$ examples, where β is the minimum slice fraction. For any scanned slab S, define

$$\widehat{\mathrm{Cov}}(S) = \frac{1}{|S|}\sum_{i \in S} H_i.$$

The reported worst-slice coverage is

$$\widehat{\mathrm{WSC}} = \min_{S \in \mathcal{S}_M} \widehat{\mathrm{Cov}}(S),$$

where $\mathcal{S}_M$ is the collection of slabs scanned over the M random projections. Equivalently, one may report worst-slice undercoverage $(\tau - \widehat{\mathrm{WSC}})_+$. The archived WSC is a descriptive minimum selected on the evaluation outcomes and covariates, including fitted $\eta_{\texttt{EE}}$. It is not an independently validated worst-group guarantee; method-specific covariates also limit direct subpopulation comparisons. By default, the number of random projections $M = 200$.

B ADDITIONAL EXPERIMENT RESULTS

B.1 RETAINED FULL NUMERICAL TABLES

The following tables retain the archived numerical estimands and evaluation rules. They are not overwritten by the common-rule component audit in Appendix B.2.

Table 3: Posterior quality on DeepSeek-V4-Flash with fixed $T = 10$. Truth labels are used only by the tilted score. NLL is the archived floored admission loss; $\widehat{p}(G)$ sums acceptable-atom mass.

Dataset	Method	NLL	$\widehat{p}(G)$	Brier	Acc@1	MRR
PopQA	FREQ-NO-MISSING	14.029	0.419	0.923	0.430	0.620
	GOOD–TURING	2.068	0.492	0.757	0.508	0.722
	PYP-EB	1.460	0.450	0.796	0.468	0.688
	BAYES-STATE	1.627	0.590	0.649	0.602	0.778
NQ-Open	FREQ-NO-MISSING	13.365	0.449	0.967	0.452	0.679
	GOOD–TURING	10.130	0.471	0.908	0.469	0.711
	PYP-EB	1.804	0.456	0.909	0.453	0.686
	BAYES-STATE	1.920	0.523	0.781	0.531	0.748

Table 4 complements the main $T = 10$ posterior table. Tables 5–6 retain every original conformal point estimate. Their legacy operational metrics are not substituted for literal-set metrics. The main delta table uses these displayed, rounded entries; it is not a new statistical estimate.

B.2 FULL MATCHED-SUPERVISION ATTRIBUTION

This study re-evaluates the complete 2×2 reference-by-tilt grid and Direct-State on the same 250 fitting, 250 calibration, and 500 test prompts over 30 shared splits. It reuses historical responses, clusters, embeddings, and fitted preliminary parameters; no new LLM calls were made. All methods use a common deterministic score jitter and order-statistic conformal quantile. Because the archived methods originally used different tie and quantile rules, these common-rule results are a component audit rather than a silent replacement of the retained main tables.

Across the eight dataset–generator–budget cells, PYP+Tilt minus PYP occupancy changes admission-mass NLL by -0.4722 (95% descriptive interval $[-0.5455, -0.3965]$). Under matched truth supervision, however, PYP+Tilt minus GT+Tilt is $+0.0120$ $[+0.0092, +0.0149]$, and PYP+Tilt minus Direct-State is $+0.0298$ $[+0.0196, +0.0396]$; positive differences are worse. Thus the matrix supports shared truth supervision, not a PYP-specific truth-estimation advantage. Direct-State uses a different derived-feature schema, so it is a strong structural control rather than a one-offset ablation.

Common fit-only groups. The raw-occupancy grouping uses only $T, K, K/T$, maximum cluster share, normalized entropy, and singleton/doubleton rates, with fit-only scaling and clustering. Equal-cell primary-target MSCE is .036066 for GT, .041349 for PYP, .011854 for GT+Tilt, .011941 for PYP+Tilt, and .012456 for Direct-State. In the frozen embedding space the corresponding values are approximately .019013, .019068, .012168, .012114, .012116. Shared truth supervision lowers these diagnostics, but no reference is uniformly best. WSC is a searched diagnostic on the same test labels, not a deployable subgroup guarantee.

B.3 DECISION-CHANNEL AND SET-SEMANTICS AUDIT

For a fixed preliminary reference, write the full score as an EE gate plus a conditional seen distribution. Gate-only preserves the learned EE mass and replaces the seen distribution by its preliminary reference; ranking-only does the converse. Each channel is recalibrated on the same calibration IDs. At the primary targets (DeepSeek .6, Mixtral .5), all 480 reference-by-unit comparisons satisfy the high-threshold condition and full versus gate-only produces identical literal sets and displayed actions. Across both target levels, 832/960 comparisons remain identical; the 128 changes occur at the lower DeepSeek target. Hence the primary fixed-budget decisions are gate-driven, but seen ranking is not generally inert.

Table 4: Posterior quality on DeepSeek-V4-Flash with fixed $T = 20$. Truth labels are used only by the tilted score. NLL is the archived floored admission loss; $\widehat{p}(G)$ sums acceptable-atom mass.

Dataset	Method	NLL	$\widehat{p}(G)$	Brier	Acc@1	MRR
PopQA	FREQ-NO-MISSING	13.491	0.427	0.899	0.442	0.614
	GOOD–TURING	1.922	0.469	0.791	0.490	0.693
	PYP-EB	1.652	0.452	0.815	0.468	0.680
	BAYES-STATE	1.710	0.579	0.669	0.591	0.770
NQ-Open	FREQ-NO-MISSING	12.939	0.456	0.947	0.458	0.676
	GOOD–TURING	10.213	0.468	0.914	0.467	0.699
	PYP-EB	1.984	0.461	0.911	0.460	0.686
	BAYES-STATE	1.962	0.515	0.792	0.519	0.739

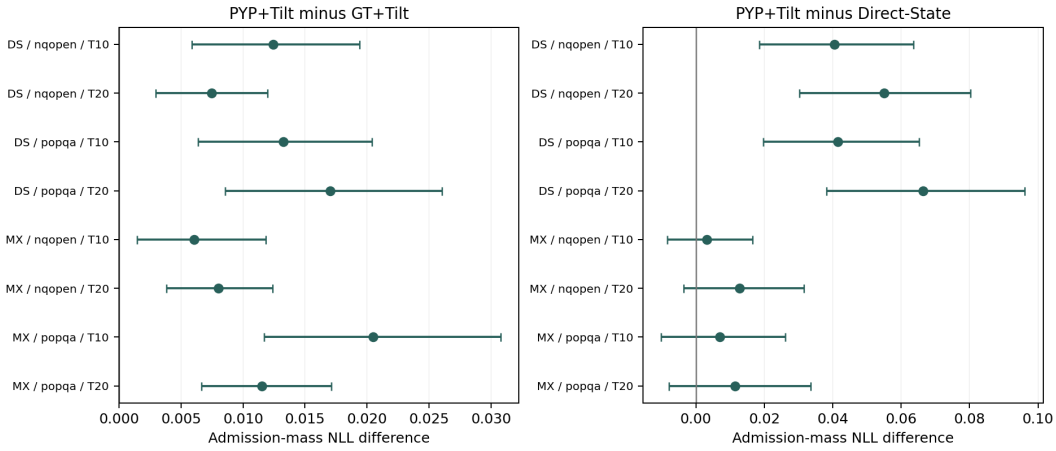

Figure 2: Cellwise admission-mass NLL contrasts. Points to the right of zero mean that PYP+Tilt has higher loss. Intervals are paired descriptive 95% question-bootstrap ranges conditional on saved fits; the 30 overlapping splits are not independent datasets.

At the primary targets, every method also includes all observed atoms in the literal threshold set, so literal seen size equals K (equal-cell mean 2.26175). The archived display policy suppresses observed atoms whenever EE passes, so displayed size can be smaller without a smaller literal candidate set. We report both quantities and do not interpret display compression as set compression.

Historical protocol and limitations. The fixed-budget tilted scores use the archived Protocol-L learner: truth-supervised preliminary selection followed by two balanced binary heads. The PYP+Tilt preliminary parameters are therefore not a truth-free species fit, and the common-rule matrix must not be relabeled as the ideal reference-fixed objective in Eq. (6). Direct-State omits reference-derived feature slots and is not a pure offset swap. Intervals condition on saved fits and reused historical questions, exclude refitting uncertainty, and are not adjusted over the many contrasts. Thirteen historical reconstructions exceeded a strict 10^{-8} numerical threshold, but all retained four-decimal summaries and saved decisions were reproduced; the maximum score discrepancy was 2.02×10^{-7}.

B.4 RETROSPECTIVE $T = 5$ REFERENCE CHECKS

The following component results reuse historical caches; no new responses were generated. They remain separate from the original 50/50-versus-25/25/50 main-table comparison.

No-tilt $T = 5$ pilot. The pilot uses DeepSeek/NQ-Open and PopQA, 1,000 prompts each, ten shared 250/250/500 splits, and target 0.60. Both arms use the same CPQ score, jitter, calibration and operational fallback; PYP is fitted from occupancy data without truth tilting. The first five

Table 5: Full retained DeepSeek-V4-Flash results, $\tau = .6$. Original operational coverage (Op.), EE inclusion, displayed size (Disp.), and group diagnostics; the z/r and WSC groupings can be method-dependent. Values are unchanged.

PopQA

T	Method	Op.	EE	Disp.	z	r	K/T	WSC
10	Seen-only	0.484	0.000	2.348	0.062	0.050	0.048	0.113
10	CPQ	0.593	0.140	1.466	0.035	0.035	0.032	0.214
10	BAYES-STATE	0.597	0.154	1.783	0.021	0.018	0.015	0.263
20	Seen-only	0.505	0.000	3.109	0.056	0.045	0.043	0.136
20	CPQ	0.595	0.110	1.940	0.038	0.031	0.028	0.229
20	BAYES-STATE	0.598	0.130	2.409	0.023	0.019	0.017	0.263

NQ-Open

T	Method	Op.	EE	Disp.	z	r	K/T	WSC
10	Seen-only	0.508	0.000	1.656	0.017	0.020	0.018	0.211
10	CPQ	0.595	0.132	1.101	0.011	0.015	0.013	0.309
10	BAYES-STATE	0.598	0.152	1.316	0.008	0.011	0.005	0.317
20	Seen-only	0.525	0.000	1.928	0.013	0.017	0.013	0.239
20	CPQ	0.596	0.117	1.223	0.010	0.015	0.011	0.316
20	BAYES-STATE	0.599	0.130	1.573	0.007	0.009	0.005	0.329

Table 6: Full retained Mixtral-8x7B-Instruct results, $\tau = .5$. Original operational coverage (Op.), EE inclusion, displayed size (Disp.), and group diagnostics; the z/r and WSC groupings can be method-dependent. Values are unchanged.

PopQA

T	Method	Op.	EE	Disp.	z	r	K/T	WSC
10	Seen-only	0.342	0.000	2.189	0.055	0.056	0.055	0.034
10	CPQ	0.502	0.182	1.215	0.024	0.047	0.044	0.091
10	BAYES-STATE	0.502	0.196	1.544	0.014	0.014	0.009	0.184
20	Seen-only	0.354	0.000	2.818	0.053	0.056	0.055	0.052
20	CPQ	0.499	0.163	1.483	0.024	0.045	0.044	0.097
20	BAYES-STATE	0.502	0.182	1.956	0.015	0.014	0.012	0.184

NQ-Open

T	Method	Op.	EE	Disp.	z	r	K/T	WSC
10	Seen-only	0.288	0.000	1.808	0.051	0.058	0.057	0.039
10	CPQ	0.501	0.246	0.824	0.022	0.057	0.056	0.184
10	BAYES-STATE	0.498	0.269	1.141	0.009	0.015	0.005	0.171
20	Seen-only	0.306	0.000	2.233	0.044	0.050	0.049	0.055
20	CPQ	0.500	0.228	0.910	0.020	0.047	0.046	0.197
20	BAYES-STATE	0.495	0.245	1.424	0.009	0.011	0.005	0.180

Table 7: Matched-supervision results for DeepSeek / NQ-Open. Literal seen size equals the observed alphabet size at these primary targets and is omitted; "Disp." is the archived fallback-collapsed display count.

T	Method	NLL	Brier	Acc@1	MRR	Op. Cov.	Disp.	EE
10	GT	2.491	0.885	0.472	0.712	0.597	1.102	0.135
	PYP occupancy	1.838	0.911	0.454	0.684	0.594	1.106	0.123
	GT+Tilt	1.787	0.769	0.530	0.748	0.598	1.321	0.154
	BAYES-STATE (PYP+Tilt)	1.800	0.771	0.530	0.749	0.598	1.318	0.154
	Direct-State	1.759	0.757	0.530	0.747	0.599	1.318	0.154
20	GT	2.607	0.897	0.466	0.699	0.600	1.225	0.123
	PYP occupancy	2.025	0.912	0.461	0.684	0.597	1.233	0.107
	GT+Tilt	1.860	0.782	0.524	0.742	0.599	1.581	0.132
	BAYES-STATE (PYP+Tilt)	1.868	0.783	0.524	0.742	0.598	1.581	0.132
	Direct-State	1.813	0.767	0.524	0.740	0.598	1.584	0.131

Table 8: Matched-supervision results for DeepSeek / PopQA. Literal seen size equals the observed alphabet size at these primary targets and is omitted; "Disp." is the archived fallback-collapsed display count.

T	Method	NLL	Brier	Acc@1	MRR	Op. Cov.	Disp.	EE
10	GT	1.929	0.756	0.509	0.722	0.604	1.429	0.151
	PYP occupancy	1.459	0.796	0.470	0.689	0.602	1.438	0.142
	GT+Tilt	1.511	0.631	0.609	0.783	0.600	1.790	0.152
	BAYES-STATE (PYP+Tilt)	1.524	0.633	0.609	0.784	0.601	1.780	0.154
	Direct-State	1.482	0.623	0.610	0.781	0.600	1.787	0.152
20	GT	2.072	0.792	0.492	0.695	0.604	1.877	0.120
	PYP occupancy	1.652	0.815	0.470	0.681	0.602	1.871	0.115
	GT+Tilt	1.593	0.651	0.599	0.775	0.603	2.389	0.132
	BAYES-STATE (PYP+Tilt)	1.610	0.653	0.600	0.776	0.602	2.402	0.130
	Direct-State	1.544	0.637	0.601	0.772	0.602	2.396	0.132

Table 9: Matched-supervision results for Mixtral / NQ-Open. Literal seen size equals the observed alphabet size at these primary targets and is omitted; "Disp." is the archived fallback-collapsed display count.

T	Method	NLL	Brier	Acc@1	MRR	Op. Cov.	Disp.	EE
10	GT	3.378	1.180	0.333	0.647	0.501	0.823	0.245
	PYP occupancy	2.381	1.220	0.282	0.594	0.499	0.821	0.241
	GT+Tilt	1.358	0.620	0.626	0.807	0.498	1.135	0.269
	BAYES-STATE (PYP+Tilt)	1.364	0.622	0.623	0.806	0.496	1.140	0.267
	Direct-State	1.361	0.622	0.623	0.804	0.498	1.134	0.270
20	GT	3.662	1.205	0.316	0.627	0.501	0.907	0.227
	PYP occupancy	2.643	1.225	0.297	0.600	0.502	0.874	0.229
	GT+Tilt	1.421	0.635	0.618	0.802	0.494	1.414	0.245
	BAYES-STATE (PYP+Tilt)	1.429	0.637	0.616	0.801	0.493	1.413	0.243
	Direct-State	1.416	0.635	0.615	0.799	0.493	1.411	0.243

Table 10: Matched-supervision results for Mixtral / PopQA. Literal seen size equals the observed alphabet size at these primary targets and is omitted; "Disp." is the archived fallback-collapsed display count.

T	Method	NLL	Brier	Acc@1	MRR	Op. Cov.	Disp.	EE
10	GT	2.616	0.972	0.391	0.662	0.498	1.231	0.179
	PYP occupancy	1.863	1.012	0.338	0.616	0.499	1.187	0.181
	GT+Tilt	1.285	0.579	0.646	0.814	0.502	1.540	0.198
	BAYES-STATE (PYP+Tilt)	1.306	0.585	0.643	0.813	0.503	1.535	0.200
	Direct-State	1.299	0.585	0.641	0.809	0.505	1.524	0.202
20	GT	2.817	1.014	0.364	0.629	0.497	1.506	0.161
	PYP occupancy	2.135	1.040	0.347	0.610	0.497	1.437	0.162
	GT+Tilt	1.307	0.578	0.645	0.812	0.502	1.951	0.183
	BAYES-STATE (PYP+Tilt)	1.318	0.582	0.642	0.811	0.501	1.959	0.182
	Direct-State	1.307	0.582	0.641	0.807	0.506	1.914	0.189

Table 11: Full versus gate-only literal-set change rates at the lower DeepSeek target $\tau = .5$. At the primary target $\tau = .6$, every corresponding rate is zero.

Dataset	Reference	T	High-threshold splits	Changed test sets (%)
NQ-Open	GT	10	12/30	3.53
NQ-Open	PYP	10	12/30	4.04
NQ-Open	GT	20	6/30	8.44
NQ-Open	PYP	20	6/30	9.57
PopQA	GT	10	24/30	1.82
PopQA	PYP	10	24/30	1.82
PopQA	GT	20	14/30	9.62
PopQA	PYP	20	14/30	10.56

assignments are a restriction of cached $T = 10$ semantic assignments, not independently constructed online $T = 5$ clusters. PopQA supports improvement on some diagnostics, while NQ-Open does not.

Table 12: Retrospective $T = 5$ no-tilt pilot. Original operational metrics and method-dependent group diagnostics; point estimates only.

Dataset	Reference	Op. Cov.	EE	Disp.	z-MSCE	r-MSCE	K/T-MSCE	WSC
NQ-Open	GT	.5950	.1674	.9580	.0100	.0198	.0151	.3115
	PYP	.5928	.1548	.9704	.0123	.0223	.0220	.3062
PopQA	GT	.5910	.1696	1.1394	.0353	.0444	.0340	.2119
	PYP	.5874	.1622	1.1544	.0351	.0375	.0364	.2365

These results motivate evaluating the PYP reference at sparse budgets; they do not establish uniform truth-coverage improvement or an online allocation advantage.

Matched supervision at $T = 5$. A separate 30-split DeepSeek/NQ-Open ablation uses the same 250/250/500 allocation as Table 1. Its retrospective $T = 5$ results are shown below. Different repetition counts and archived evaluations mean this table and the preceding pilot are not pooled into one estimator comparison. The close supervised rows are evidence for shared truth-aware learning,

Table 13: Matched-supervision $T = 5$ ablation, DeepSeek/NQ-Open, 30 splits.

Method	Op. Cov.	EE	Disp.	z-MSCE	r-MSCE	K/T-MSCE	WSC
GT	.601	.160	1.302	.008	.038	.032	.265
PYP	.601	.160	1.298	.008	.033	.032	.272
GT+Tilt	.609	.194	1.459	.007	.017	.007	.316
BAYES-STATE	.609	.196	1.452	.007	.018	.007	.316
Direct-State	.607	.191	1.461	.007	.018	.007	.309

not evidence that a PYP offset is essential. The comparison does not establish equivalence from nonsignificance, or isolate Bayesian estimation from every generic supervised alternative.

Sampling-law tests and sequential limits. Cold-start reference quality should first be assessed on held-out *generation* outcomes, such as whether the next draw is a new species. Its Bernoulli probability is sampling missing mass, not truth-absence probability. A second observable event is that the next draw is new and the following draw repeats it; under PYP its probability is

$$\frac{(\theta + \sigma K_t)(1 - \sigma)}{(t + \theta)(t + \theta + 1)}.$$

Appendix C.2 now supplies truth-free fits and information-matched sampling controls. Its next-response and fixed-alphabet count targets differ from the truth-labeled tables above; retrospective semantic slices are not treated as online partitions.

In the stable-process special case $\theta = 0$, the above probability is $\sigma(1 - \sigma)K_t/[t(t + 1)]$. With shared $0 < \sigma < 1$, its ordering equals the parameter-free index $K_t/[t(t + 1)]$. A free stopping threshold

absorbs the common scale as well. The subsequent audit found identical stop vectors in 240/240 checked cases; apparent gains over raw doubleton-based GT therefore cannot by themselves be attributed to learning the discount. Stronger smoothing controls also limit that interpretation. We retain the analytic cold-start reference, not a claim of superior adaptive allocation.

Interpretation and remaining validation. The reported comparisons separate a sampling reference from shared truth-aware learning. They do not establish that more elaborate priors or posterior computation improve conditional coverage. Sampling-law fit, truth supervision, and the set constructor should be assessed under separate matched controls. These reused historical cohorts support component analysis, not independent confirmation or automatic transfer to new generators.

B.5 Results on PopQA and NQ-Open with standard deviation

In the following tables, "Tilt" denotes BAYES-STATE. Columns share the same dataset, generator, and target; entries retain the original means and split standard deviations.

Table 14: DeepSeek-V4-Flash results for PopQA at target $\tau = 0.5$. Operational coverage/display metrics; mean $\pm$ split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.484 ± 0.015	0.492 ± 0.030	0.498 ± 0.031	0.499 ± 0.026	0.495 ± 0.031	0.502 ± 0.035
Displayed size	2.348 ± 0.056	2.088 ± 0.278	2.182 ± 0.194	2.711 ± 0.650	2.114 ± 0.552	2.570 ± 0.576
EE fraction	0.000 ± 0.000	0.013 ± 0.014	0.022 ± 0.027	0.000 ± 0.000	0.002 ± 0.006	0.010 ± 0.021
MSCE_z	0.048 ± 0.006	0.046 ± 0.007	0.044 ± 0.010	0.047 ± 0.005	0.048 ± 0.006	0.047 ± 0.008
MSCE_r	0.037 ± 0.006	0.036 ± 0.007	0.039 ± 0.009	0.038 ± 0.008	0.041 ± 0.008	0.042 ± 0.009
$\mathrm{MSCE}_{K/T}$	0.035 ± 0.006	0.032 ± 0.009	0.032 ± 0.009	0.037 ± 0.008	0.039 ± 0.008	0.037 ± 0.009

Table 15: DeepSeek-V4-Flash results for PopQA at target $\tau = 0.6$. Operational coverage/display metrics; mean $\pm$ split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.484 ± 0.015	0.593 ± 0.030	0.597 ± 0.041	0.505 ± 0.016	0.595 ± 0.031	0.598 ± 0.037
Displayed size	2.348 ± 0.056	1.466 ± 0.129	1.783 ± 0.172	3.109 ± 0.099	1.940 ± 0.184	2.409 ± 0.253
EE fraction	0.000 ± 0.000	0.140 ± 0.028	0.154 ± 0.047	0.000 ± 0.000	0.110 ± 0.025	0.130 ± 0.042
MSCE_z	0.062 ± 0.007	0.035 ± 0.007	0.021 ± 0.008	0.056 ± 0.007	0.038 ± 0.006	0.023 ± 0.008
MSCE_r	0.050 ± 0.007	0.035 ± 0.006	0.018 ± 0.008	0.045 ± 0.007	0.031 ± 0.004	0.019 ± 0.008
$\mathrm{MSCE}_{K/T}$	0.048 ± 0.007	0.032 ± 0.005	0.015 ± 0.008	0.043 ± 0.007	0.028 ± 0.004	0.017 ± 0.008

Table 16: DeepSeek-V4-Flash results for NQ-Open at target $\tau = 0.5$. Operational coverage/display metrics; mean $\pm$ split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.505 ± 0.020	0.503 ± 0.025	0.501 ± 0.031	0.506 ± 0.024	0.502 ± 0.023	0.504 ± 0.032
Displayed size	1.600 ± 0.157	1.472 ± 0.165	1.443 ± 0.157	1.433 ± 0.263	1.356 ± 0.197	1.524 ± 0.250
EE fraction	0.000 ± 0.000	0.004 ± 0.009	0.011 ± 0.020	0.000 ± 0.000	0.000 ± 0.000	0.004 ± 0.009
MSCE_z	0.008 ± 0.003	0.008 ± 0.003	0.009 ± 0.004	0.008 ± 0.003	0.008 ± 0.003	0.009 ± 0.003
MSCE_r	0.013 ± 0.005	0.015 ± 0.006	0.017 ± 0.006	0.018 ± 0.005	0.020 ± 0.005	0.018 ± 0.006
$\mathrm{MSCE}_{K/T}$	0.011 ± 0.005	0.012 ± 0.005	0.012 ± 0.004	0.013 ± 0.005	0.015 ± 0.005	0.013 ± 0.005

Table 17: DeepSeek-V4-Flash results for NQ-Open at target $\tau = 0.6$. Operational coverage/display metrics; mean ± split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.508 ± 0.013	0.595 ± 0.029	0.598 ± 0.038	0.525 ± 0.012	0.596 ± 0.025	0.599 ± 0.032
Displayed size	1.656 ± 0.034	1.101 ± 0.071	1.316 ± 0.113	1.928 ± 0.052	1.223 ± 0.079	1.573 ± 0.120
EE fraction	0.000 ± 0.000	0.132 ± 0.033	0.152 ± 0.051	0.000 ± 0.000	0.117 ± 0.027	0.130 ± 0.045
MSCE_z	0.017 ± 0.004	0.011 ± 0.003	0.008 ± 0.003	0.013 ± 0.003	0.010 ± 0.003	0.007 ± 0.002
MSCE_r	0.020 ± 0.004	0.015 ± 0.002	0.011 ± 0.004	0.017 ± 0.004	0.015 ± 0.002	0.009 ± 0.004
$\mathrm{MSCE}_{K/T}$	0.018 ± 0.004	0.013 ± 0.002	0.005 ± 0.003	0.013 ± 0.004	0.011 ± 0.002	0.005 ± 0.003

Table 18: Mixtral-8x7B-Instruct results for PopQA at target $\tau = 0.5$. Operational coverage/display metrics; mean ± split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.342 ± 0.016	0.502 ± 0.035	0.502 ± 0.036	0.354 ± 0.016	0.499 ± 0.029	0.502 ± 0.038
Displayed size	2.189 ± 0.046	1.215 ± 0.097	1.544 ± 0.140	2.818 ± 0.078	1.483 ± 0.114	1.956 ± 0.224
EE fraction	0.000 ± 0.000	0.182 ± 0.026	0.196 ± 0.038	0.000 ± 0.000	0.163 ± 0.022	0.182 ± 0.040
MSCE_z	0.055 ± 0.007	0.024 ± 0.005	0.014 ± 0.005	0.053 ± 0.007	0.024 ± 0.006	0.015 ± 0.006
MSCE_r	0.056 ± 0.006	0.047 ± 0.006	0.014 ± 0.005	0.056 ± 0.007	0.045 ± 0.005	0.014 ± 0.007
$\mathrm{MSCE}_{K/T}$	0.055 ± 0.006	0.044 ± 0.007	0.009 ± 0.005	0.055 ± 0.007	0.044 ± 0.005	0.012 ± 0.006

Table 19: Mixtral-8x7B-Instruct results for PopQA at target $\tau = 0.6$. Operational coverage/display metrics; mean ± split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.342 ± 0.016	0.603 ± 0.029	0.598 ± 0.035	0.354 ± 0.016	0.600 ± 0.028	0.601 ± 0.037
Displayed size	2.189 ± 0.046	0.866 ± 0.052	1.192 ± 0.116	2.818 ± 0.078	0.973 ± 0.057	1.445 ± 0.183
EE fraction	0.000 ± 0.000	0.302 ± 0.023	0.320 ± 0.043	0.000 ± 0.000	0.286 ± 0.026	0.310 ± 0.048
MSCE_z	0.097 ± 0.009	0.022 ± 0.006	0.012 ± 0.005	0.092 ± 0.009	0.020 ± 0.006	0.013 ± 0.005
MSCE_r	0.097 ± 0.008	0.040 ± 0.004	0.017 ± 0.006	0.095 ± 0.009	0.048 ± 0.008	0.011 ± 0.004
$\mathrm{MSCE}_{K/T}$	0.097 ± 0.009	0.038 ± 0.004	0.006 ± 0.004	0.094 ± 0.009	0.047 ± 0.007	0.008 ± 0.004

Table 20: Mixtral-8x7B-Instruct results for NQ-Open at target $\tau = 0.5$. Operational coverage/display metrics; mean ± split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.288 ± 0.015	0.501 ± 0.031	0.498 ± 0.041	0.306 ± 0.015	0.500 ± 0.032	0.495 ± 0.035
Displayed size	1.808 ± 0.053	0.824 ± 0.055	1.141 ± 0.117	2.233 ± 0.088	0.910 ± 0.066	1.424 ± 0.173
EE fraction	0.000 ± 0.000	0.246 ± 0.030	0.269 ± 0.043	0.000 ± 0.000	0.228 ± 0.033	0.245 ± 0.037
MSCE_z	0.051 ± 0.006	0.022 ± 0.007	0.009 ± 0.004	0.044 ± 0.006	0.020 ± 0.005	0.009 ± 0.004
MSCE_r	0.058 ± 0.006	0.057 ± 0.008	0.015 ± 0.006	0.050 ± 0.005	0.047 ± 0.007	0.011 ± 0.004
$\mathrm{MSCE}_{K/T}$	0.057 ± 0.006	0.056 ± 0.008	0.005 ± 0.004	0.049 ± 0.005	0.046 ± 0.007	0.005 ± 0.003

Table 21: Mixtral-8x7B-Instruct results for NQ-Open at target $\tau = 0.6$. Operational coverage/display metrics; mean ± split std. See Appendix A.8 for grouping caveats.

	$T = 10$			$T = 20$		
Metric	Seen-only	CPQ	Tilt	Seen-only	CPQ	Tilt
Op. coverage	0.288 ± 0.015	0.611 ± 0.033	0.596 ± 0.038	0.306 ± 0.015	0.609 ± 0.029	0.602 ± 0.032
Displayed size	1.808 ± 0.053	0.611 ± 0.039	0.881 ± 0.112	2.233 ± 0.088	0.627 ± 0.034	1.046 ± 0.142
EE fraction	0.000 ± 0.000	0.399 ± 0.039	0.398 ± 0.042	0.000 ± 0.000	0.375 ± 0.034	0.389 ± 0.034
MSCE_z	0.103 ± 0.009	0.020 ± 0.005	0.010 ± 0.004	0.093 ± 0.009	0.019 ± 0.005	0.010 ± 0.003
MSCE_r	0.110 ± 0.009	0.059 ± 0.009	0.030 ± 0.009	0.099 ± 0.008	0.075 ± 0.010	0.021 ± 0.007
$\mathrm{MSCE}_{K/T}$	0.109 ± 0.009	0.059 ± 0.009	0.007 ± 0.005	0.098 ± 0.008	0.075 ± 0.010	0.006 ± 0.004

C Cold-start sampling uncertainty and its limits

This appendix evaluates the PYP component on its own statistical target: the generator's sampling law. It does not fit a new truth model or replace the fixed-budget truth-score results. The experiment is historical development evidence from existing caches, with no new model calls or correctness labels. It separates M, the number of historical occupancy-training questions, from t, the online query prefix; each historical question has 20 responses. Neither resource is the truth-label fitting budget in the main experiment.

C.1 Conditional residual-mass and future-count laws

For prompt x, assume exchangeable semantic responses from $G_x \sim \mathrm{PY}(\sigma, \theta, H_x)$ with diffuse base H_x. The discount and strength obey $0 \le \sigma < 1$ and $\theta > -\sigma$, excluding $(\sigma, \theta) = (0, 0)$. The fitted experiment restricts $\theta \ge 0$, with explicit Dirichlet process (DP; $\sigma = 0$) and stable-process ($\theta = 0$) boundaries. For a partition with counts $n_1, \ldots, n_K$ summing to t, the exchangeable partition probability function (EPPF) is

$$p_{\sigma,\theta}(n_{1:K}) = \frac{\prod_{j=1}^{K-1}(\theta + j\sigma)}{(\theta+1)_{t-1}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1}, \tag{9}$$

where $(a)_m$ is the rising factorial. Hyperparameters maximize the summed log EPPF of M disjoint historical questions, then remain fixed at test time. This is a plug-in empirical-Bayes reference, not integration over hyperparameter uncertainty.

The underlying stick-breaking construction is

$$G_x = \sum_{j \ge 1} p_j \delta_{\zeta_j}, \qquad p_j = V_j \prod_{\ell < j}(1 - V_\ell), \qquad V_j \sim \mathrm{Beta}(1-\sigma, \theta + j\sigma), \quad \zeta_j \sim H_x.$$

Let $U_t = G_x(\mathcal{A}_t^c)$ be residual sampling mass outside the *currently* observed alphabet. Conditional on the prefix and fixed (σ, θ),

$$\big(G_x(a_1), \ldots, G_x(a_K), U_t\big) \sim \mathrm{Dirichlet}(n_1 - \sigma, \ldots, n_K - \sigma, \theta + \sigma K). \tag{10}$$

Thus $U_t \sim \mathrm{Beta}(a, b)$ with $a = \theta + \sigma K$ and $b = t - \sigma K$. Its mean gives the one-step novelty probability in Eq. (2), but the full law also represents residual-mass uncertainty.

For the next h responses define

$$Y_{t,h} = \sum_{j=1}^{h} \mathbf{1}\{\kappa_x(Z_{t+j}) \notin \mathcal{A}_t\}.$$

Repeated responses from a species first seen within the future window each count in $Y_{t,h}$. This is *not* the number of distinct future species, and it is not a truth-recovery event.

Proposition C.1 (Residual-mass propagation)**.** *Under the stated model, conditional on the prefix and fixed hyperparameters,*

$$P(Y_{t,h} = y) = \binom{h}{y} \frac{\mathrm{B}(a + y, b + h - y)}{\mathrm{B}(a, b)}, \qquad y = 0, \ldots, h.$$

The integrated beta-binomial law and its plug-in $\mathrm{Binomial}(h, q)$ *comparator,* $q = a/(a+b)$*, have the same mean* hq*. Their variance ratio is* $1 + (h-1)/(t+\theta+1)$*.*

Proof. Given G_x, each future draw lies outside the fixed alphabet with probability U_t, so $Y_{t,h} \mid U_t \sim \mathrm{Binomial}(h, U_t)$. Integrating against the beta density gives the displayed mass function. The law of total variance yields

$$\mathrm{Var}(Y) = h\mathbb{E}[U_t(1-U_t)] + h^2\,\mathrm{Var}(U_t) = hq(1-q)\frac{a+b+h}{a+b+1}.$$

Since $a+b = t+\theta$, division by the plug-in variance proves the ratio. □

For $\theta = 0$, $h = 5$, and $t = 2, 3, 5$, the ratios are $7/3$, 2, and $5/3$. At $h = 1$ the predictive laws coincide; for fixed h the variance excess vanishes as t grows. This is an exact conditional calculation, not an appeal to a general Bernstein–von Mises theorem at a particular finite budget.

If P denotes the integrated model law and Q its plug-in law, then

$$\mathbb{E}_P[-\log Q(Y)] - \mathbb{E}_P[-\log P(Y)] = \mathrm{KL}(P\|Q) \geq 0.$$

The unnormalized ranked probability score (RPS) is $\mathrm{RPS}(F, y) = \sum_{j=0}^{h-1}(F(j) - \mathbf{1}\{y \leq j\})^2$, its expected excess under Q is $\sum_j (F_P(j) - F_Q(j))^2$. These are proper-score identities *under the assumed integrated law*. They do not guarantee empirical superiority under misspecification, conditional conformal coverage, or sharper intervals. The mechanism also applies to DP and non-Bayesian beta-binomial forecasts; it is not unique to PYP.

C.2 Data, information budgets, and controls

Six historical cells combine NQ-Open, PopQA, and TriviaQA with DeepSeek and Mixtral. Each supplies the first 20 responses for 1,000 questions: 120,000 responses, but only 3,000 dataset-qualified question identities because generators share questions. The four NQ-Open/PopQA cells are primary; TriviaQA is a prespecified historical sensitivity, not untouched confirmation. Later Mixtral/TriviaQA responses are unused. Historical aliases lack immutable model-revision evidence.

The primary alphabet independently normalizes each response without accessing future responses or reference answers. The semantic sensitivity assigns each response to its first matching cached representative and never changes earlier assignments. Nontransitive cached judgments make this alphabet path-dependent, so exchangeability is a modeling approximation. One unresolved Mixtral/NQ-Open semantic trajectory is excluded before splitting; its exact-normalized trajectory remains included. No new semantic judgments are obtained.

For repeats $0, \ldots, 9$, source-order IDs are shuffled using the repeat seed. The first 250 form a nested training pool, the next 250 are reserved and unused, and the remaining 499/500 form the test set. Fits use $M \in \{5, 10, 25, 50, 250\}$ questions, or $20M$ historical responses, with zero correctness labels. Prefixes are $t \in \{2, 3, 5, 10, 15\}$ and windows $h \in \{5, 10\}$ satisfying $t + h \leq 20$. The main aggregate uses $M = 5, 10, 25$, $t = 2, 3, 5$, $h = 5$. The 600 fitted units are 6 cells $\times 2$ alphabets $\times 5$ training budgets $\times 10$ repeats.

One-step controls are fitted PYP, fitted DP, the existing fixed PYP $(\sigma, \theta) = (.3, 2)$, raw GT N_1/t, stabilized GT $(N_1 + .5)/(t+1)$, a pooled-by-t Jeffreys novelty rate, and occupancy logistic cross-validation (CV). The latter uses question-grouped three-fold CV over $C \in \{.1, 1, 10\}$, fold-specific scaling, and all 19 transitions per training question. Its validation questions contribute no fitting transitions. Remaining mass for frequency-based controls is allocated to seen atoms in proportion to counts.

Count forecasts compare PYP integration to its identical-mean binomial law, with the same paired comparison for DP. Strong controls include stabilized-GT binomial, occupancy-CV binomial, occupancy beta-binomial, and pooled beta-binomial. Occupancy beta-binomial concentration is fitted to whole-pipeline, nested out-of-fold novelty forecasts, not in-sample means. Concentration bounds $[.02, 10000]$ and single-class Jeffreys fallbacks are fixed. Raw GT zero support and infinite losses are retained rather than epsilon-clipped.

All reported uncertainty intervals first average repeated held-out appearances within question and configurations within question, then weight cells equally. A 2,000-draw dataset-qualified question bootstrap shares weights across generators. Intervals condition on saved fits, exclude refitting uncertainty, and are unadjusted descriptive intervals on reused historical data. The ten panels overlap; neither panels nor table cells are independent replications.

C.3 Results and their limits

Table 22 separates one-step estimation from multi-step uncertainty propagation. The primary practical one-step screen requires categorical NLL difference at most $-.01$, interval upper bound below zero, and non-worsening categorical Brier. PYP passes against DP, but not against stabilized GT or occupancy CV. No universal point-estimation advantage is established.

Table 22: Primary exact-normalized comparisons. Negative differences favor the left method. $M = 5, 10, 25$, $t = 2, 3, 5$; the count task uses $h = 5$. Intervals are descriptive 95% ranges conditional on fitted models.

Comparison	Target	Δ	95% interval
PYP − DP	Category NLL	-0.1344	$[-0.1407, -0.1277]$
PYP − DP	Category Brier	-0.1017	$[-0.1066, -0.0967]$
PYP − stabilized GT	Category NLL	-0.0049	$[-0.0141, +0.0039]$
PYP − stabilized GT	Category Brier	+0.0006	$[-0.0044, +0.0056]$
PYP − occupancy CV	Category NLL	-0.0072	$[-0.0129, -0.0017]$
PYP − occupancy CV	Category Brier	+0.0026	$[+0.0001, +0.0052]$
Integrated − mean plug-in	Count NLL	-0.4174	$[-0.4351, -0.3993]$
Integrated − mean plug-in	RPS	-0.0857	$[-0.0891, -0.0824]$
Integrated − occupancy beta-binomial (BB)	Count NLL	-0.0713	$[-0.0880, -0.0549]$
Integrated − occupancy BB	RPS	+0.0132	$[+0.0043, +0.0221]$

The integrated PYP law improves count NLL and RPS over the mean plug-in in all six dataset/generator cells and both alphabets on the specified small-budget aggregate. Predictive-mean absolute error (MAE) is unchanged by construction; the largest saved numerical mean discrepancy was 3.64×10^{-14}. This isolates distributional uncertainty propagation rather than a better point forecast.

Against occupancy beta-binomial, the advantage is narrower. At $M = 5$, both aggregate scores favor PYP; by $M = 25$, both favor the occupancy control. Figure 3 and Table 23 retain all historical-data budgets. At $M = 5$, the RPS direction favors PYP in only 5/10 overlapping exact-alphabet panels (7/10 semantic panels). These aggregates do not establish stable refitting gains. Pooling $M = 5, 10, 25$ gives better PYP NLL but worse RPS.

Table 23: PYP integrated − occupancy beta-binomial at every historical training budget. $t = 2, 3, 5$, $h = 5$; four core cells. The sign reversal at larger M is retained.

Alphabet	M	Δ NLL	95% interval	Δ RPS	95% interval
Exact	5	-0.2079	$[-0.2301, -0.1861]$	-0.0158	$[-0.0253, -0.0064]$
Exact	10	-0.0476	$[-0.0638, -0.0317]$	+0.0138	$[+0.0049, +0.0227]$
Exact	25	+0.0415	$[+0.0272, +0.0551]$	+0.0417	$[+0.0322, +0.0510]$
Exact	50	+0.0630	$[+0.0489, +0.0763]$	+0.0477	$[+0.0382, +0.0568]$
Exact	250	+0.0807	$[+0.0677, +0.0927]$	+0.0560	$[+0.0471, +0.0645]$
Semantic	5	-0.3819	$[-0.4175, -0.3449]$	-0.0144	$[-0.0208, -0.0080]$
Semantic	10	-0.1178	$[-0.1348, -0.1004]$	+0.0186	$[+0.0128, +0.0240]$
Semantic	25	+0.0232	$[+0.0122, +0.0341]$	+0.0283	$[+0.0220, +0.0347]$
Semantic	50	+0.0464	$[+0.0355, +0.0574]$	+0.0332	$[+0.0263, +0.0401]$
Semantic	250	+0.0623	$[+0.0521, +0.0725]$	+0.0389	$[+0.0324, +0.0457]$

Sensitivity and numerical boundaries. Table 25 includes both alphabets, the larger window, later prefixes, and TriviaQA. The future-count target remains fixed-alphabet response counts in

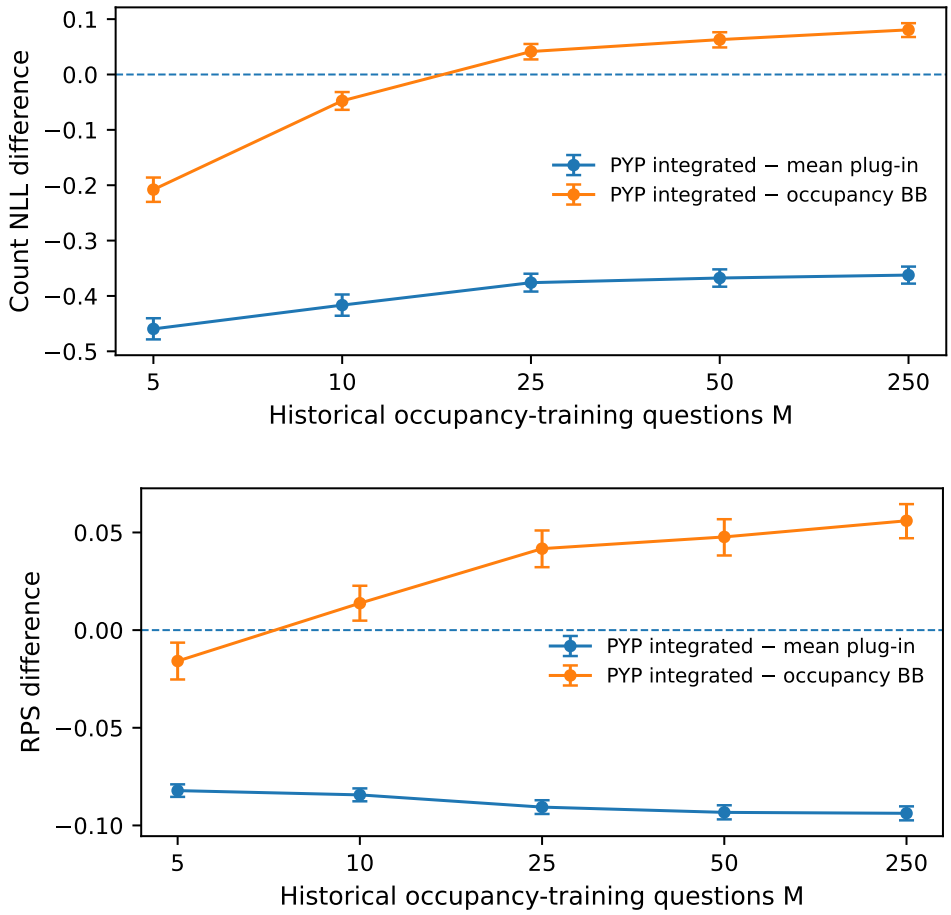

Figure 3: Historical-data boundary of the cold-start claim, exact-normalized alphabet. Error bars are descriptive 95% question-bootstrap intervals conditional on fitted models. Integrating the PYP residual mass improves on its mean plug-in, but does not uniformly improve on the information-matched occupancy beta-binomial. Lower is better.

Table 24: Count prediction and nominal 90% equal-tail predictive intervals, primary exact alphabet. Width is in response counts. PYP coverage above nominal accompanies wider intervals; it is not evidence of sharper calibrated UQ.

Method	NLL	RPS	Mean MAE	Cov. (%)	Width
PYP integrated	1.3702	0.7413	1.2239	96.54	3.880
PYP mean-binomial	1.7876	0.8270	1.2239	86.55	2.814
DP integrated	1.8813	1.1017	1.6379	88.70	3.989
DP mean-binomial	2.4652	1.1900	1.6379	72.19	3.080
Stabilized GT binomial	1.8117	0.7778	1.1227	85.62	2.381
Occupancy binomial	1.8141	0.7685	1.0911	83.26	2.240
Occupancy beta-binomial	1.4416	0.7280	1.0911	93.93	3.395
Pooled beta-binomial	1.6633	0.9816	1.5440	95.20	4.406

every row. Integration does not overturn prior negative results for *distinct*-species forecasts or truth recovery. TriviaQA favors occupancy beta-binomial on both primary proper scores in the exact alphabet. Semantic results additionally depend on the approximate exchangeability of the first-matching alphabet. All 600 selected PYP and 600 selected DP fits returned optimizer success; unsuccessful candidate starts are preserved in the source report. PYP strength is near zero in 536/600

Table 25: Prespecified sensitivities: PYP integrated minus each comparator. Small-prefix rows average $M = 5, 10, 25$, $t = 2, 3, 5$ and $h = 5$ unless otherwise stated. Complete intervals and cell-level tables are retained in the machine-readable supplement.

Aggregate	Alphabet	Comparator	Δ NLL	Δ RPS
Core, small prefixes	Exact	Plug-in	-0.4174	-0.0857
Core, small prefixes	Exact	Occupancy BB	-0.0713	+0.0132
Core, small prefixes	Semantic	Plug-in	-0.3166	-0.0484
Core, small prefixes	Semantic	Occupancy BB	-0.1588	+0.0108
Core, $h = 10$	Exact	Plug-in	-1.1865	-0.2678
Core, $h = 10$	Exact	Occupancy BB	-0.0158	+0.0429
Core, $h = 10$	Semantic	Plug-in	-0.8987	-0.1692
Core, $h = 10$	Semantic	Occupancy BB	-0.1944	+0.0307
Core, $t = 10, 15$	Exact	Plug-in	-0.0775	-0.0109
Core, $t = 10, 15$	Exact	Occupancy BB	-0.0756	+0.0032
Core, $t = 10, 15$	Semantic	Plug-in	-0.0623	-0.0031
Core, $t = 10, 15$	Semantic	Occupancy BB	-0.1741	-0.0046
TriviaQA, small prefixes	Exact	Plug-in	-0.6083	-0.0570
TriviaQA, small prefixes	Exact	Occupancy BB	+0.1353	+0.0447
TriviaQA, small prefixes	Semantic	Plug-in	-0.5321	-0.0218
TriviaQA, small prefixes	Semantic	Occupancy BB	+0.1858	+0.0054

selected fits. Concentration fits also frequently hit the prespecified bounds, especially at $M = 5$; optimizer success does not remove restrictions from the prespecified parameter domains.

What can be concluded. The evidence supports a parsimonious, coherent mechanism for propagating residual-mass uncertainty under short online prefixes, with a bounded scarce-history development advantage over the tested flexible estimator. It does not establish general Bayesian superiority, universal label efficiency, more accurate truth identification, or lower conformal cost. The count intervals are posterior-predictive model outputs, not distribution-free conformal intervals. Historical reuse, overlapping training panels, alias-version uncertainty, and semantic approximation limit external generalization. No new adaptive policy was executed.

D Proofs and Additional Theoretical Results

D.1 Reference fixing and identification

Proof of Proposition 3.1. For each state, $b'_a e^{g'_a} = e^{c(B)} b_a e^{g_a}$, so the common factor cancels on normalization. Conversely, for a fixed positive b and positive categorical law $\eta^\star$, equality $\eta^\star_a = b_a e^{g_a}/Z$ gives $g_a = \log(\eta^\star_a/b_a) + \log Z$. The condition $g_{\mathtt{EE}} = 0$ fixes $\log Z = -\log(\eta^\star_{\mathtt{EE}}/b_{\mathtt{EE}})$, proving the formula and uniqueness of the relative function. □

The proposition concerns functions on a fixed alphabet, not unique weights of a parametrized learner. Zero truth probabilities require extended log ratios or limits; a zero reference cannot represent positive truth mass through a finite correction. Adding a small fixed positive reference mass is a support convention, not evidence that the corresponding event has that probability. Its value and normalization must be identical in a matched comparison.

When an admission set G has several elements, its summed score does not identify the within-set probabilities. For example, if the only observed admission set is $\{a_1, a_2\}$, every categorical vector with $\eta_1 + \eta_2 = 1$ minimizes its negative log admission mass. No unique law of a selected truth atom follows without a specified selection or coarsening model. The valid claim is therefore a learned admission score plus the intersection coverage result below. This resolves the interpretation without replacing retained surrogate-fit results by unrun categorical-likelihood results.

For the cold-start study, the species reference is fitted by the marginal occupancy likelihood, independently of correctness labels. For the archived full score, supervised preliminary tuning is instead part of the score learner. These are distinct fitting procedures. Reference-free and GT-based learners may use the same truth supervision. Component attribution therefore requires matched labels and reference swaps, rather than a Bayesian label.

D.2 Proof of Theorem 3.2

Proof. Condition on the fitting split. Since the fitted posterior and score are frozen, the calibration scores $R_1, \dots, R_{n_{\mathrm{cal}}}$ and the test score

$$R_{n+1} = \widehat{S}_{n+1}(A^\star_{n+1})$$

are exchangeable. Moreover,

$$A^\star_{n+1} \in \widehat{C}_T(X_{n+1}) \quad \Longleftrightarrow \quad R_{n+1} \le \widehat{q}.$$

By the standard split-conformal rank argument,

$$\mathbb{P}(R_{n+1} \le R_{(k)} \mid \mathcal{D}_{\mathrm{fit}}) \ge \frac{k}{n_{\mathrm{cal}} + 1} \ge 1 - \alpha.$$

Taking expectation over the fitting split gives the claim. □

D.3 Set-valued marginal validity

In the main text, we state marginal validity for the single-answer case with a collapsed truth atom $A^\star_i$. We now give the corresponding set-valued version. For each example i, finite querying and semantic clustering induce a local alphabet $\mathcal{Y}_i = \mathcal{A}_i \cup \{\mathtt{EE}\}$. Let $G_i \subseteq \mathcal{Y}_i$ be the nonempty set of acceptable truth atoms. If no observed semantic atom is acceptable, then $G_i = \{\mathtt{EE}\}$.

After empirical-Bayes fitting, the score $\widehat{S}_i(a)$ is frozen. Define the set-valued calibration score

$$R_i = \min_{a \in G_i} \widehat{S}_i(a), \qquad i \in \mathcal{I}_{\mathrm{cal}}.$$

Let $n_{\mathrm{cal}} = |\mathcal{I}_{\mathrm{cal}}|$, $k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$, and $\widehat{q} = R_{(k)}$, with the convention $R_{(n_{\mathrm{cal}}+1)} = \infty$. For a new prompt, define

$$\widehat{C}(X_{n+1}) = \{a \in \mathcal{Y}_{n+1} : \widehat{S}_{n+1}(a) \le \widehat{q}\}.$$

Theorem D.1 (Set-valued marginal validity)**.** *Assume that, conditional on the fitting split, the calibration examples and the test example are exchangeable after including the LLM query randomness, semantic clustering, local alphabet, and admission set G_i. Then*

$$\mathbb{P}\Big(\widehat{C}(X_{n+1}) \cap G_{n+1} \neq \emptyset\Big) \geq 1 - \alpha.$$

Proof. Condition on the fitting split. Since the fitted posterior and score are frozen, the set-valued scores

$$R_i = \min_{a \in G_i} \widehat{S}_i(a), \qquad i \in \mathcal{I}_{\text{cal}} \cup \{n+1\},$$

are exchangeable. Moreover,

$$\widehat{C}(X_{n+1}) \cap G_{n+1} \neq \emptyset \quad \Longleftrightarrow \quad \min_{a \in G_{n+1}} \widehat{S}_{n+1}(a) \leq \widehat{q} \quad \Longleftrightarrow \quad R_{n+1} \leq \widehat{q}.$$

By the standard split-conformal rank argument,

$$\mathbb{P}(R_{n+1} \leq R_{(k)} \mid \mathcal{D}_{\text{fit}}) \geq \frac{k}{n_{\text{cal}} + 1} \geq 1 - \alpha.$$

Taking expectation over the fitting split gives the result. □

D.4 Proof of Theorem 3.3

Proof. Condition on all randomness used to construct $\widehat{\eta}$ and the threshold q, so that both are fixed for a fresh test state B. Let

$$\varepsilon(B) = \|\widehat{\eta}_B - \eta^\star_B\|_\infty.$$

For every atom $a \in \mathcal{Y}_B$,

$$|S_{\widehat{\eta}}(B, a) - S_{\eta^\star}(B, a)| \leq \varepsilon(B).$$

This also holds for EE, since the same EE shift is applied to both scores. Hence, if an atom belongs to exactly one of $C_{\widehat{\eta},q}(B)$ and $C_{\eta^\star,q}(B)$, then its oracle score must be within $\varepsilon(B)$ of the threshold:

$$a \in C_{\widehat{\eta},q}(B) \triangle C_{\eta^\star,q}(B) \quad \Longrightarrow \quad |S_{\eta^\star}(B, a) - q| \leq \varepsilon(B).$$

Therefore,

$$C_{\widehat{\eta},q}(B) \triangle C_{\eta^\star,q}(B) \subseteq \{a \in \mathcal{Y}_B : |S_{\eta^\star}(B, a) - q| \leq \varepsilon(B)\}.$$

Taking $\eta^\star_B$-mass and using the boundary-mass condition gives

$$|\eta^\star_B(C_{\widehat{\eta},q}) - \eta^\star_B(C_{\eta^\star,q})| \leq \Gamma_B(q, \varepsilon(B)) \leq L\varepsilon(B).$$

Let

$$U(B) = \eta^\star_B(C_{\widehat{\eta},q}) - \tau, \qquad V(B) = \eta^\star_B(C_{\eta^\star,q}) - \tau.$$

By the reverse triangle inequality in L^2,

$$\sqrt{\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q)} = \|U\|_{L^2} \leq \|V\|_{L^2} + \|U - V\|_{L^2} \leq \sqrt{\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q)} + L\sqrt{\mathbb{E}[\varepsilon(B)^2]}.$$

Finally, writing TV for total-variation distance,

$$\varepsilon(B) \leq \|\widehat{\eta}_B - \eta^\star_B\|_1 = 2\,\mathrm{TV}(\eta^\star_B, \widehat{\eta}_B).$$

By Pinsker's inequality,

$$\varepsilon(B)^2 \leq 4\,\mathrm{TV}^2(\eta^\star_B, \widehat{\eta}_B) \leq 2\,\mathrm{KL}(\eta^\star_B \,\|\, \widehat{\eta}_B).$$

Taking expectations yields

$$\mathbb{E}[\varepsilon(B)^2] \leq 2\mathcal{K}(\widehat{\eta}).$$

Combining the displays proves

$$\sqrt{\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q)} \leq \sqrt{\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q)} + \sqrt{2}\,L\,\sqrt{\mathcal{K}(\widehat{\eta})}.$$

□

D.5 FINITE-SAMPLE GROUP-MSCE CERTIFICATE

The exact state-conditional MSCE is not directly observable from one truth label per prompt. In experiments, we therefore evaluate its group-based approximation. We now give a finite-sample certificate for this empirical MSCE.

Let $h(B) \in [M]$ be a fixed grouping rule, independent of the evaluation labels. For a score-induced set $C_{\eta,q}$, define the group-conditional coverage

$$\pi_g(\eta, q) = \mathbb{P}(A^\star \in C_{\eta,q}(B) \mid h(B) = g), \qquad p_g = \mathbb{P}(h(B) = g),$$

and the population group-MSCE

$$\mathrm{MSCE}_h(\eta, q) = \sum_{g=1}^{M} p_g \big(\pi_g(\eta, q) - \tau\big)^2.$$

Given an independent evaluation sample $\{(B_j, A_j^\star)\}_{j=1}^m$, let $n_g = \sum_{j=1}^m \mathbf{1}\{h(B_j) = g\}$ and

$$\widehat{\pi}_g(\eta, q) = \frac{1}{n_g} \sum_{j:h(B_j)=g} \mathbf{1}\{A_j^\star \in C_{\eta,q}(B_j)\}$$

for groups with $n_g > 0$. The empirical group-MSCE is

$$\widehat{\mathrm{MSCE}}_{h,m}(\eta, q) = \sum_{g:n_g>0} \frac{n_g}{m} \big(\widehat{\pi}_g(\eta, q) - \tau\big)^2.$$

Theorem D.2 (Finite-sample group-MSCE certificate)**.** *Fix η, q, and h. With probability at least $1 - \delta$ over the evaluation sample,*

$$\left|\widehat{\mathrm{MSCE}}_{h,m}(\eta, q) - \mathrm{MSCE}_h(\eta, q)\right| \le 2 \max_{g:n_g>0} \sqrt{\frac{\log(4M/\delta)}{2n_g}} + \sqrt{\frac{\log(4/\delta)}{2m}}.$$

Proof. Conditional on the group counts, Hoeffding's inequality gives, uniformly over all nonempty groups, with probability at least $1 - \delta/2$,

$$|\widehat{\pi}_g(\eta, q) - \pi_g(\eta, q)| \le \sqrt{\frac{\log(4M/\delta)}{2n_g}}.$$

Since $u \mapsto (u - \tau)^2$ is 2-Lipschitz on $[0, 1]$, this implies

$$\left|\sum_{g:n_g>0} \frac{n_g}{m} \Big[(\widehat{\pi}_g(\eta, q) - \tau)^2 - (\pi_g(\eta, q) - \tau)^2\Big]\right| \le 2 \max_{g:n_g>0} \sqrt{\frac{\log(4M/\delta)}{2n_g}}.$$

It remains to control the random group proportions. Let $Z = h(B)$ and $z_g = (\pi_g(\eta, q) - \tau)^2 \in [0, 1]$. Then $\sum_g (n_g/m) z_g$ is the empirical mean of the bounded random variable $z_{h(B)}$, whose expectation is $\mathrm{MSCE}_h(\eta, q)$. Hoeffding's inequality gives, with probability at least $1 - \delta/2$,

$$\left|\sum_g \frac{n_g}{m} z_g - \sum_g p_g z_g\right| \le \sqrt{\frac{\log(4/\delta)}{2m}}.$$

Combining the two events by a union bound proves the claim. □

D.6 SCORE-LEVEL STABILITY OF HIGH-PROBABILITY SETS

The posterior-to-MSCE theorem (Theorem 3.3) treats $\widehat{C}$ as an arbitrary posterior-induced set rule. The next result uses the specific structure of the high-probability score and shows that posterior estimation error affects the set only through the oracle probability mass near the score decision boundary.

Proposition D.3 (Score-level stability of high-probability sets)**.** *For any posterior η_B on $\mathcal{Y}_B$, define*

$$S_\eta(B,a) = \begin{cases} 1-\eta_B(a), & a \in \mathcal{A}(B), \\ 2-\eta_B(\texttt{EE}), & a = \texttt{EE}. \end{cases}$$

Let $S^\star = S_{\eta^\star}$ and $S_{\widehat{\eta}}$ be the scores induced by $\eta_B^\star$ and $\widehat{\eta}_B$. For a threshold q, define

$$C_q^\star(B) = \{a \in \mathcal{Y}_B : S^\star(B,a) \le q\}, \qquad \widehat{C}_q(B) = \{a \in \mathcal{Y}_B : S_{\widehat{\eta}}(B,a) \le q\}.$$

Let $\varepsilon(B) = \|\widehat{\eta}_B - \eta_B^\star\|_\infty$ and define the oracle boundary mass

$$\Gamma_B(q,r) = \eta_B^\star(\{a \in \mathcal{Y}_B : |S^\star(B,a) - q| \le r\}).$$

Then, for every state B,

$$\left|\eta_B^\star(\widehat{C}_q) - \eta_B^\star(C_q^\star)\right| \le \Gamma_B(q, \varepsilon(B)).$$

Consequently,

$$\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q) \le 2\,\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q) + 2\mathbb{E}\big[\Gamma_B(q,\varepsilon(B))^2\big],$$

where

$$\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q) = \mathbb{E}\left[\Big(\eta_B^\star(\widehat{C}_q) - \tau\Big)^2\right], \qquad \mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q) = \mathbb{E}\Big[\big(\eta_B^\star(C_q^\star) - \tau\big)^2\Big].$$

If $\Gamma_B(q,r) \le L_B r$ for the relevant range of r, then

$$\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q) \le 2\,\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q) + 2\mathbb{E}[L_B^2 \varepsilon(B)^2].$$

In particular, if $L_B \le L$ almost surely, then

$$\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q) \le 2\,\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q) + 4L^2 \mathcal{K}(\widehat{\eta}).$$

Proof. For every atom $a \in \mathcal{Y}_B$,

$$|S_{\widehat{\eta}}(B,a) - S^\star(B,a)| \le \|\widehat{\eta}_B - \eta_B^\star\|_\infty = \varepsilon(B).$$

This holds for both seen atoms and EE, since the same EE shift is applied to both scores. Therefore, if an atom belongs to exactly one of $\widehat{C}_q(B)$ and $C_q^\star(B)$, then its oracle score must lie within $\varepsilon(B)$ of the threshold:

$$a \in \widehat{C}_q(B) \triangle C_q^\star(B) \quad \Longrightarrow \quad |S^\star(B,a) - q| \le \varepsilon(B).$$

Thus

$$\widehat{C}_q(B) \triangle C_q^\star(B) \subseteq \{a \in \mathcal{Y}_B : |S^\star(B,a) - q| \le \varepsilon(B)\}.$$

Taking $\eta_B^\star$-mass gives

$$\left|\eta_B^\star(\widehat{C}_q) - \eta_B^\star(C_q^\star)\right| \le \eta_B^\star(\widehat{C}_q \triangle C_q^\star) \le \Gamma_B(q, \varepsilon(B)).$$

Let $u(B) = \eta_B^\star(\widehat{C}_q)$ and $v(B) = \eta_B^\star(C_q^\star)$. Since $|u(B) - v(B)| \le \Gamma_B(q,\varepsilon(B))$,

$$|u(B) - \tau| \le |v(B) - \tau| + \Gamma_B(q,\varepsilon(B)).$$

Squaring, using $(a+b)^2 \le 2a^2 + 2b^2$, and taking expectations yields

$$\mathrm{MSCE}_{\mathrm{B}}(\widehat{\eta}, q) \le 2\,\mathrm{MSCE}_{\mathrm{B}}(\eta^\star, q) + 2\mathbb{E}[\Gamma_B(q,\varepsilon(B))^2].$$

The margin-bound statement follows from $\Gamma_B(q,r) \le L_B r$.

Finally,

$$\varepsilon(B) \le \|\widehat{\eta}_B - \eta_B^\star\|_1 = 2\,\mathrm{TV}(\eta_B^\star, \widehat{\eta}_B).$$

By Pinsker's inequality,

$$\varepsilon(B)^2 \le 4\,\mathrm{TV}^2(\eta_B^\star, \widehat{\eta}_B) \le 2\,\mathrm{KL}(\eta_B^\star \| \widehat{\eta}_B).$$

If $L_B \le L$, then

$$2\mathbb{E}[L_B^2 \varepsilon(B)^2] \le 4L^2 \mathcal{K}(\widehat{\eta}),$$

which proves the final bound. □

D.7 STATE-CONDITIONAL MSCE CONTROLS PROMPT-CONDITIONAL MSCE

Let

$$H = \mathbf{1}\{A^\star \in \widehat{C}(B)\}$$

be the coverage indicator for a fresh test example. In the set-valued QA case, replace it by

$$H = \mathbf{1}\{G \cap \widehat{C}(B) \neq \emptyset\}.$$

For any conditioning variable W, define

$$\pi_W(W) = \mathbb{E}[H \mid W], \qquad \mathrm{MSCE}_W = \mathbb{E}[(\pi_W(W) - \tau)^2].$$

Proposition D.4 (Coarsening contracts MSCE)**.** *Let W_0 and W_1 be two conditioning variables such that*

$$\sigma(W_0) \subseteq \sigma(W_1).$$

Then

$$\pi_{W_0}(W_0) = \mathbb{E}[\pi_{W_1}(W_1) \mid W_0],$$

and

$$\mathrm{MSCE}_{W_1} = \mathrm{MSCE}_{W_0} + \mathbb{E}\Big[\left(\pi_{W_1}(W_1) - \pi_{W_0}(W_0)\right)^2\Big].$$

In particular,

$$\mathrm{MSCE}_{W_0} \leq \mathrm{MSCE}_{W_1}.$$

Taking $W_0 = X$ and $W_1 = B$ gives

$$\mathrm{MSCE}_{\mathrm{X}} \leq \mathrm{MSCE}_{\mathrm{B}}.$$

Proof. By the tower property and the assumption $\sigma(W_0) \subseteq \sigma(W_1)$,

$$\pi_{W_0}(W_0) = \mathbb{E}[H \mid W_0] = \mathbb{E}[\mathbb{E}[H \mid W_1] \mid W_0] = \mathbb{E}[\pi_{W_1}(W_1) \mid W_0].$$

Therefore

$$\pi_{W_0}(W_0) - \tau = \mathbb{E}[\pi_{W_1}(W_1) - \tau \mid W_0].$$

Jensen's inequality gives

$$\left(\pi_{W_0}(W_0) - \tau\right)^2 \leq \mathbb{E}[(\pi_{W_1}(W_1) - \tau)^2 \mid W_0].$$

Taking expectations yields

$$\mathrm{MSCE}_{W_0} \leq \mathrm{MSCE}_{W_1}.$$

The stronger identity follows by orthogonal decomposition. Write

$$\pi_{W_1}(W_1) - \tau = (\pi_{W_0}(W_0) - \tau) + \left(\pi_{W_1}(W_1) - \pi_{W_0}(W_0)\right).$$

Squaring and taking expectations,

$$\begin{aligned}\mathrm{MSCE}_{W_1} = \mathrm{MSCE}_{W_0} &+ \mathbb{E}[(\pi_{W_1}(W_1) - \pi_{W_0}(W_0))^2] \\ &+ 2\mathbb{E}[(\pi_{W_0}(W_0) - \tau)(\pi_{W_1}(W_1) - \pi_{W_0}(W_0))].\end{aligned}$$

The cross term is zero because

$$\begin{aligned}&\mathbb{E}[(\pi_{W_0}(W_0) - \tau)(\pi_{W_1}(W_1) - \pi_{W_0}(W_0))] \\ &= \mathbb{E}[(\pi_{W_0}(W_0) - \tau)\mathbb{E}[\pi_{W_1}(W_1) - \pi_{W_0}(W_0) \mid W_0]] \\ &= 0.\end{aligned}$$

This proves the identity and the inequality. □

D.8 SET-VALUED POSTERIOR-TO-MSCE TRANSFER

We now give the set-valued analogue. Let $G \subseteq \mathcal{Y}_B$ be the acceptable atom set returned by the admission rule. For a posterior η_B and threshold q, define

$$H_{\eta,q} = \mathbb{1}\{G \cap C_{\eta,q}(B) \neq \emptyset\}, \qquad \pi^{\mathrm{set}}_{\eta,q}(B) = \mathbb{P}(H_{\eta,q} = 1 \mid B).$$

The set-valued state-conditional MSCE is

$$\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\eta, q) = \mathbb{E}\Big[\big(\pi^{\mathrm{set}}_{\eta,q}(B) - \tau\big)^2\Big].$$

For threshold q and tolerance $\xi \geq 0$, define the set-valued oracle boundary mass

$$\Gamma^{\mathrm{set}}_B(q, \xi) = \mathbb{P}(G \cap \{a \in \mathcal{Y}_B : |S_{\eta^\star}(B, a) - q| \leq \xi\} \neq \emptyset \,|\, B)\,.$$

Proposition D.5 (Set-valued posterior-to-MSCE transfer)**.** *Condition on the fitted score and threshold q, with a fresh test observation retaining the stated conditional truth law. Suppose that there exists $L < \infty$ such that $\Gamma^{\mathrm{set}}_B(q, \xi) \leq L\xi$ for all relevant B and ξ. Then*

$$\sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\widehat{\eta}, q)} \leq \sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\eta^\star, q)} + L\sqrt{\mathbb{E}[\|\widehat{\eta}_B - \eta^\star_B\|^2_\infty]}.$$

If, in addition, $\mathbb{E}[\|\widehat{\eta}_B - \eta^\star_B\|^2_\infty] \leq 2\mathcal{K}(\widehat{\eta})$, then

$$\sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\widehat{\eta}, q)} \leq \sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\eta^\star, q)} + \sqrt{2}\, L\, \sqrt{\mathcal{K}(\widehat{\eta})}.$$

Proof. Condition on all randomness used to construct $\widehat{\eta}$ and q. Let

$$\varepsilon(B) = \|\widehat{\eta}_B - \eta^\star_B\|_\infty.$$

As in the single-answer proof,

$$C_{\widehat{\eta},q}(B) \triangle C_{\eta^\star,q}(B) \subseteq \{a \in \mathcal{Y}_B : |S_{\eta^\star}(B, a) - q| \leq \varepsilon(B)\}.$$

If the two set-valued coverage indicators differ, then the acceptable set must intersect the symmetric difference:

$$H_{\widehat{\eta},q} \neq H_{\eta^\star,q} \quad \Longrightarrow \quad G \cap \big(C_{\widehat{\eta},q}(B) \triangle C_{\eta^\star,q}(B)\big) \neq \emptyset.$$

Therefore,

$$|\pi^{\mathrm{set}}_{\widehat{\eta},q}(B) - \pi^{\mathrm{set}}_{\eta^\star,q}(B)| \leq \Gamma^{\mathrm{set}}_B(q, \varepsilon(B)) \leq L\varepsilon(B).$$

Let

$$U(B) = \pi^{\mathrm{set}}_{\widehat{\eta},q}(B) - \tau, \qquad V(B) = \pi^{\mathrm{set}}_{\eta^\star,q}(B) - \tau.$$

The reverse triangle inequality in L^2 gives

$$\sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\widehat{\eta}, q)} = \|U\|_{L^2} \leq \|V\|_{L^2} + \|U - V\|_{L^2} \leq \sqrt{\mathrm{MSCE}^{\mathrm{set}}_{\mathrm{B}}(\eta^\star, q)} + L\sqrt{\mathbb{E}[\varepsilon(B)^2]}.$$

This proves the first claim. The second follows immediately from the additional bound

$$\mathbb{E}[\varepsilon(B)^2] \leq 2\mathcal{K}(\widehat{\eta}).$$

□