A. Extended Background

A.1. Conformal Factuality

Let $X \in \mathcal{X}$ be a prompt and let $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$ denote a fixed generator over completions. At inference time, we repeatedly draw candidates $Y \sim \pi(\cdot \mid X)$ and seek a prediction set that contains *at least one correct answer* with high probability:

$$\mathbb{P}\Big(\exists y \in \widehat{C}_\alpha(X_{N+1}) : A(X_{N+1}, y) = 1\Big) \geq 1 - \alpha, \tag{A.1}$$

where α is the target error level and $A(X, y) \in \{0, 1\}$ indicates whether candidate y is correct for prompt X. To achieve this guarantee, each sampled candidate is evaluated with a verifier score $V : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. In this paper, smaller verifier scores are better, so a calibrated acceptance rule amounts to choosing a threshold and retaining all candidates with score below it.

A.2. Inductive Conformal Prediction

Split conformal prediction transforms the outputs of a black-box model into valid prediction sets using a held-out calibration set [17, 20]. Given calibration pairs $\{(X_i, Y_i)\}_{i=1}^N$ and a nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, one computes calibration scores $V_i = s(X_i, Y_i)$, sorts them, and forms the empirical quantile

$$\widehat{Q}_{1-\alpha} := V_{(\lceil (N+1)(1-\alpha) \rceil)}. \tag{A.2}$$

The corresponding split-conformal prediction set is

$$\widehat{C}_{1-\alpha}(X_{N+1}) := \{y \in \mathcal{Y} : s(X_{N+1}, y) \leq \widehat{Q}_{1-\alpha}\}, \tag{A.3}$$

which satisfies marginal coverage

$$\mathbb{P}\Big(Y_{N+1} \in \widehat{C}_{1-\alpha}(X_{N+1})\Big) \geq 1 - \alpha \tag{A.4}$$

under exchangeability. In the conformal-factuality setting, this corresponds to using one global acceptance threshold for all prompts.

A.3. Conditional Conformal Prediction

Marginal coverage holds only on average over the prompt distribution. For LLMs, this can hide severe heterogeneity: easy prompts may be over-covered while hard prompts are under-covered. The ideal pointwise conditional guarantee

$$\mathbb{P}\Big(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1} = x\Big) \geq 1 - \alpha \tag{A.5}$$

is impossible to achieve exactly in finite samples without strong assumptions [6, 19].

Following Gibbs et al. [7], one can rewrite exact conditional coverage as an infinite family of weighted marginal constraints:

$$\mathbb{P}\Big(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1}\Big) = 1 - \alpha \iff \mathbb{E}\Big[f(X_{N+1})\big(\mathbb{1}\{Y_{N+1} \in \widehat{C}(X_{N+1})\} - (1-\alpha)\big)\Big] = 0 \tag{A.6}$$

for all measurable f. Their relaxation replaces the class of all measurable functions with a chosen function class $\mathcal{F}$:

$$\mathbb{E}\Big[f(X_{N+1})\big(\mathbb{1}\{Y_{N+1} \in \widehat{C}(X_{N+1})\} - (1-\alpha)\big)\Big] = 0, \qquad \text{for all } f \in \mathcal{F}. \tag{A.7}$$

Taking $\mathcal{F} = \{1\}$ recovers marginal conformal prediction, while $\mathcal{F} = \{\Phi(\cdot)^\top \beta : \beta \in \mathbb{R}^d\}$ yields a finite-dimensional feature-conditional target.

For this linear class, Gibbs et al. define an augmented quantile-regression estimator using the pinball loss

$$\rho_{1-\alpha}(u) = u\big((1-\alpha) - \mathbb{1}\{u < 0\}\big). \tag{A.8}$$

Given calibration scores $\{(X_i, S_i)\}_{i=1}^N$ and a fresh candidate score S, the augmented fit is

$$\widehat{g}_S := \operatorname*{arg\,min}_{g \in \mathcal{F}} \frac{1}{N+1} \sum_{i=1}^{N} \rho_{1-\alpha}\big(S_i - g(X_i)\big) \; + \; \frac{1}{N+1} \rho_{1-\alpha}\big(S - g(X_{N+1})\big). \tag{A.9}$$

The resulting prediction rule keeps labels whose score does not exceed the fitted value at the same score:

$$\widehat{C}(X_{N+1}) := \left\{y : S(X_{N+1}, y) \leq \widehat{g}_{S(X_{N+1}, y)}(X_{N+1})\right\}. \tag{A.10}$$

Theorem A.1 (Gibbs et al. [7], Theorem 2)**.** *Let $\mathcal{F} = \{\Phi(\cdot)^\top \beta : \beta \in \mathbb{R}^d\}$ be a linear class over the basis $\Phi : \mathcal{X} \to \mathbb{R}^d$. Then for any non-negative $f \in \mathcal{F}$ with $\mathbb{E}[f(X)] > 0$, the prediction rule above satisfies*

$$\mathbb{P}_f\Big(Y_{N+1} \in \widehat{C}(X_{N+1})\Big) \geq 1 - \alpha. \tag{A.11}$$

Our method is a conformal-factuality instantiation of this framework, with the latent success score $S(X)$ replacing the standard label-wise nonconformity score and a fixed-point construction used to obtain the deployed threshold.

B. Proof of Results from the Main Paper

B.1. Proof of Theorem 4.1

Proof. Recall the success score

$$S(X) \;:=\; \inf\{\lambda \in [0,1] :\; \ell_\lambda(X) = 0\} \;=\; \inf\{V(X,y) : y \in C(X),\; A(X,y) = 1\},$$

so that for any threshold $\lambda(X) \in [0,1]$,

$$\big\{\exists\, y \in \widehat{C}_\alpha(X) : A(X,y) = 1\big\} \quad\Longleftrightarrow\quad \big\{S(X) \leq \lambda(X)\big\}. \tag{B.1}$$

Thus, if we show that

$$\mathbb{P}_f\big(S(X_{N+1}) \leq \widehat{\lambda}_\alpha(X_{N+1})\big) \;\geq\; 1 - \alpha \tag{B.2}$$

for every non-negative $f \in \mathcal{F}$ with $\mathbb{E}[f(X)] > 0$, the result follows immediately from (B.1).

Throughout, for any such f we define the f-reweighted probability of an event E by

$$\mathbb{P}_f(E) := \frac{\mathbb{E}\big[f(X)\,\mathbb{1}\{(X,S) \in E\}\big]}{\mathbb{E}[f(X)]}.$$

Let $\tau := 1 - \alpha$ and recall the pinball loss

$$\rho_\tau(u) \;=\; u\big(\tau - \mathbb{1}\{u < 0\}\big).$$

Let β_S be the augmented quantile-regression minimizer from Eq. (B.5) for the realized calibration set and the fresh pair (X,S), and define $g(S) := \Phi(X)^\top \beta_S$. Then for any nonnegative $f \in \mathcal{F}$ with $\mathbb{E}[f(X)] > 0$,

$$\mathbb{E}\big[f(X)\,\mathbb{1}\{S \leq g(S)\}\big] \;\geq\; (1-\alpha)\,\mathbb{E}[f(X)].$$

Fix $\gamma \in \mathbb{R}^d$ with $f(X) = \Phi(X)^\top \gamma \geq 0$ for all X and consider

$$\psi_N(\varepsilon) = \frac{1}{N+1}\sum_{i=1}^{N} \rho_{1-\alpha}\big(S_i - \Phi(X_i)^\top(\beta_S + \varepsilon\gamma)\big) + \frac{1}{N+1}\rho_{1-\alpha}\big(S - \Phi(X)^\top(\beta_S + \varepsilon\gamma)\big).$$

By convexity and optimality of β_S, $\psi_N'(0^+) \geq 0$. This gives

$$\psi_N'(0^+) = -\frac{1}{N+1}\left[\sum_{i=1}^{N}\big((1-\alpha) - \mathbb{1}\{S_i \leq \Phi(X_i)^\top\beta_S\}\big)f(X_i) + \big((1-\alpha) - \mathbb{1}\{S \leq \Phi(X)^\top\beta_S\}\big)f(X)\right] \geq 0.$$

Rearranging,

$$\frac{1}{N+1}\sum_{i=1}^{N}\mathbb{1}\{S_i \leq \Phi(X_i)^\top\beta_S\}f(X_i) + \frac{1}{N+1}\mathbb{1}\{S \leq \Phi(X)^\top\beta_S\}f(X) \;\geq\; (1-\alpha)\,\frac{1}{N+1}\sum_{i=1}^{N} f(X_i) + (1-\alpha)\,\frac{1}{N+1}f(X).$$

Now take expectation over all $N+1$ exchangeable draws. By exchangeability, each of the $N+1$ summands on each side has the same distribution, so

$$\mathbb{E}\big[f(X)\,\mathbb{1}\{S \leq \Phi(X)^\top\beta_S\}\big] \;\geq\; (1-\alpha)\,\mathbb{E}[f(X)].$$

Since $g(S) = \Phi(X)^\top \beta_S$, the claim follows.

Recall the CFC threshold

$$\widehat{\lambda}_\alpha(X) \;:=\; \sup\{\, t \in [0,1] :\; t \le g(t)\,\}, \qquad g(t) = \Phi(X)^\top \beta_t.$$

By definition of the supremum, for any realized $S \in [0,1]$ we have

$$\mathbb{1}\{S \le g(S)\} \;\le\; \mathbb{1}\{S \le \widehat{\lambda}_\alpha(X)\}. \tag{B.3}$$

Multiplying (B.3) by $f(X) \ge 0$ and taking expectations,

$$\mathbb{E}\big[f(X)\,\mathbb{1}\{S \le \widehat{\lambda}_\alpha(X)\}\big] \;\ge\; \mathbb{E}\big[f(X)\,\mathbb{1}\{S \le g(S)\}\big] \;\ge\; (1-\alpha)\,\mathbb{E}[f(X)].$$

Dividing by $\mathbb{E}[f(X)] > 0$ gives

$$\mathbb{P}_f\big(S \le \widehat{\lambda}_\alpha(X)\big) \;\ge\; 1-\alpha.$$

By definition of the success score and the prediction set,

$$\{\, S(X) \le \widehat{\lambda}_\alpha(X)\,\} \quad\Longleftrightarrow\quad \{\, \exists\, y \in \widehat{C}_\alpha(X) : A(X,y) = 1\,\}.$$

Therefore,

$$\mathbb{P}_f\Big(\exists\, y \in \widehat{C}_\alpha(X) : A(X,y) = 1\Big) \;=\; \mathbb{P}_f\Big(S(X) \le \widehat{\lambda}_\alpha(X)\Big) \;\ge\; 1-\alpha,$$

as claimed. □

B.2. Proof of Theorem 4.2

Proof. Write the calibration set as

$$\mathcal{D}_{\text{cal}} = \{(X_i, S_i)\}_{i=1}^N, \qquad (X_{N+1}, S_{N+1}) \sim \text{i.i.d. as } (X_i, S_i),$$

and recall the success indicator

$$Z(x,s) \;:=\; \mathbb{1}\big\{s \le \widehat{\lambda}_\alpha(x)\big\},$$

so that $Z(X,S) = 1$ iff there exists a correct candidate in $\widehat{C}_\alpha(X)$.

Define

$$Q(\mathcal{D}_{\text{cal}}) \;:=\; \mathbb{P}\big(Z(X,S) = 1 \mid \mathcal{D}_{\text{cal}}\big) \;=\; \mathbb{E}\big[Z(X,S) \mid \mathcal{D}_{\text{cal}}\big].$$

By Theorem 4.1 applied with $f \equiv 1$ and the law of total expectation,

$$\mathbb{P}\big(Z(X,S) = 1\big) = \mathbb{E}_{\mathcal{D}_{\text{cal}}}\, Q(\mathcal{D}_{\text{cal}}) \;\ge\; 1-\alpha.$$

Thus

$$\mathbb{E}_{\mathcal{D}_{\text{cal}}}\big[Q(\mathcal{D}_{\text{cal}})\big] \;\ge\; 1-\alpha. \tag{B.4}$$

We now show that $Q(\mathcal{D}_{\text{cal}})$ is Lipschitz in the calibration set, with sensitivity of order $1/N$ to replacing one calibration pair.

Let $\mathcal{D} = \{(X_i, S_i)\}_{i=1}^N$ and $\mathcal{D}' = \{(X_i', S_i')\}_{i=1}^N$ differ only in the k-th pair. For a fixed test prompt x and a candidate success score $s \in [0,1]$, consider the ridge-regularized augmented quantile regression objective

$$\beta \;\mapsto\; \frac{1}{N+1}\sum_{i=1}^N \rho_{1-\alpha}\big(S_i - \Phi(X_i)^\top \beta\big) + \frac{1}{N+1}\rho_{1-\alpha}\big(s - \Phi(x)^\top \beta\big) + \frac{\lambda}{2}\|\beta\|_2^2. \tag{B.5}$$

Let $\beta_s(\mathcal{D}, x)$ and $\beta_s(\mathcal{D}', x)$ denote the unique minimizers of (B.5) with $\mathcal{D}$ and $\mathcal{D}'$ respectively, and define

$$g_{\mathcal{D}}(x,s) \;:=\; \Phi(x)^\top \beta_s(\mathcal{D}, x), \qquad g_{\mathcal{D}'}(x,s) \;:=\; \Phi(x)^\top \beta_s(\mathcal{D}', x).$$

By Assumption (1), $\|\Phi(X_i)\|_2 \le R$ almost surely, and the pinball loss is 1-Lipschitz. Hence each sample loss is R-Lipschitz in β. By Assumption (2), adding the ridge term makes the objective (B.5) strongly convex. Standard stability results for regularized ERM imply that replacing one data point perturbs the minimizer by at most

$$\left\|\beta_s(\mathcal{D}, x) - \beta_s(\mathcal{D}', x)\right\|_2 \;\le\; \frac{C_1 R}{\lambda N}$$

for some universal constant $C_1 > 0$, uniformly over $s \in [0,1]$ and x.

Using Assumption (1) again,

$$\left|g_{\mathcal{D}}(x,s) - g_{\mathcal{D}'}(x,s)\right| \le \|\Phi(x)\|_2 \left\|\beta_s(\mathcal{D}, x) - \beta_s(\mathcal{D}', x)\right\|_2 \le \frac{C_1 R^2}{\lambda N}.$$

Thus there is a constant $C_2 > 0$ such that

$$\sup_{x \in \mathcal{X}} \sup_{s \in [0,1]} \left|g_{\mathcal{D}}(x,s) - g_{\mathcal{D}'}(x,s)\right| \;\le\; \frac{C_2}{N}. \tag{B.6}$$

The CFC threshold at x is

$$\widehat{\lambda}_\alpha^{(\mathcal{D})}(x) \;:=\; \sup\{\, s \in [0,1] :\; s \le g_{\mathcal{D}}(x,s)\,\},$$

and similarly for $\widehat{\lambda}_\alpha^{(\mathcal{D}')}(x)$. Under the fixed-point construction in Eq. (4.3), this threshold is a Lipschitz functional of $s \mapsto g_{\mathcal{D}}(x,s)$ in the uniform norm: there exists a constant $C_3 > 0$ such that

$$\sup_{x \in \mathcal{X}} \left|\widehat{\lambda}_\alpha^{(\mathcal{D})}(x) - \widehat{\lambda}_\alpha^{(\mathcal{D}')}(x)\right| \;\le\; \frac{C_3}{N}. \tag{B.7}$$

For a fresh test pair (X, S),

$$Q(\mathcal{D}) = \mathbb{P}\big(S \le \widehat{\lambda}_\alpha^{(\mathcal{D})}(X) \mid \mathcal{D}\big) = \mathbb{E}\Big[F_{S|X}\big(\widehat{\lambda}_\alpha^{(\mathcal{D})}(X)\big) \;\Big|\; \mathcal{D}\Big],$$

where $F_{S|X=x}(\cdot)$ is the conditional CDF of S given $X = x$.

By Assumption (3), for each x the map $t \mapsto F_{S|X=x}(t)$ is L-Lipschitz on $[0,1]$. Combining this with (B.7) yields

$$\left|Q(\mathcal{D}) - Q(\mathcal{D}')\right| \le L \sup_{x \in \mathcal{X}} \left|\widehat{\lambda}_\alpha^{(\mathcal{D})}(x) - \widehat{\lambda}_\alpha^{(\mathcal{D}')}(x)\right| \le \frac{C_4}{N},$$

where $C_4 := LC_3$. Thus $Q(\mathcal{D}_{\text{cal}})$ satisfies bounded differences with constants $c_i = C_4/N$ for $i = 1, \dots, N$.

By McDiarmid's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\big(Q(\mathcal{D}_{\text{cal}}) \le \mathbb{E}_{\mathcal{D}_{\text{cal}}}\big[Q(\mathcal{D}_{\text{cal}})\big] - \varepsilon\big) \le \exp\left(-\frac{2N\varepsilon^2}{C_4^2}\right).$$

Given $\delta \in (0,1)$, set

$$\varepsilon_N(\delta) := \frac{C_4}{\sqrt{2N}} \sqrt{\log \frac{1}{\delta}} = O\left(\sqrt{\frac{\log(1/\delta)}{N}}\right).$$

Then with probability at least $1 - \delta$ over $\mathcal{D}_{\text{cal}}$,

$$Q(\mathcal{D}_{\text{cal}}) \;\ge\; \mathbb{E}_{\mathcal{D}_{\text{cal}}}\big[Q(\mathcal{D}_{\text{cal}})\big] - \varepsilon_N(\delta). \tag{B.8}$$

Combining (B.8) with (B.4) yields

$$Q(\mathcal{D}_{\text{cal}}) \;\ge\; 1 - \alpha - \varepsilon_N(\delta)$$

with probability at least $1 - \delta$, i.e.

$$\mathbb{P}(Z(X,S) = 1 \mid \mathcal{D}_{\text{cal}}) \;\ge\; 1 - \alpha - \varepsilon_N(\delta),$$

Finally, we have $\alpha_{\text{eff}} = \max\{0, \alpha - \varepsilon_N(\delta)\} = \alpha - \varepsilon_N(\delta)$ (the slack is a small term) from algorithm 2, then,

$$\mathbb{P}\Big(S \le \widehat{\lambda}_{\alpha_{\text{eff}}}(X) \mid \mathcal{D}_{\text{cal}}\Big) \ge 1 - \alpha_{\text{eff}} - \varepsilon_N(\delta) = 1 - \alpha,$$

□

B.3. Proof of Proposition 4.3

Proof. Let

$$u_0 := 1 - \alpha, \qquad u_{\bar{\lambda}}(t) := F_t(\bar{\lambda}).$$

Since $\mathbb{P}(S \le \bar{\lambda}) = 1 - \alpha$, we have

$$\mathbb{E}\big[u_{\bar{\lambda}}(T)\big] = \mathbb{E}\big[F_T(\bar{\lambda})\big] = 1 - \alpha = u_0.$$

By Assumption 2, the map $t \mapsto u_{\bar{\lambda}}(t)$ is nonincreasing.

Now,

$$\mathbb{E}\big[G_X(\lambda^\star(X))\big] = \mathbb{E}\big[G_T(q_\alpha(T))\big] = \mathbb{E}\big[C_T(u_0)\big],$$

because $q_\alpha(t) = F_t^{-1}(u_0)$, and also

$$\mathbb{E}\big[G_X(\bar{\lambda})\big] = \mathbb{E}\big[G_T(\bar{\lambda})\big] = \mathbb{E}\big[C_T(u_{\bar{\lambda}}(T))\big].$$

Set

$$a(t) := u_{\bar{\lambda}}(t) - u_0.$$

Then $a(t)$ is nonincreasing in t and

$$\mathbb{E}[a(T)] = 0.$$

By convexity of $C_t(\cdot)$, for each t,

$$C_t(u_{\bar{\lambda}}(t)) \ge C_t(u_0) + \partial_u C_t(u_0)\,\big(u_{\bar{\lambda}}(t) - u_0\big).$$

Hence

$$C_t(u_{\bar{\lambda}}(t)) - C_t(u_0) \ge m(t)\,a(t), \qquad m(t) := \partial_u C_t(u_0).$$

By Assumption 4, $m(t)$ is nonincreasing in t. Since both $m(t)$ and $a(t)$ are nonincreasing, Chebyshev's rearrangement inequality gives

$$\mathbb{E}[m(T)a(T)] \ge \mathbb{E}[m(T)]\,\mathbb{E}[a(T)] = 0.$$

Taking expectations in the previous convexity bound yields

$$\mathbb{E}\big[C_T(u_{\bar{\lambda}}(T))\big] - \mathbb{E}\big[C_T(u_0)\big] \ge \mathbb{E}[m(T)a(T)] \ge 0.$$

Therefore,

$$\mathbb{E}\big[G_X(\bar{\lambda})\big] = \mathbb{E}\big[C_T(u_{\bar{\lambda}}(T))\big] \ge \mathbb{E}\big[C_T(u_0)\big] = \mathbb{E}\big[G_X(\lambda^\star(X))\big].$$

For strictness, if $\mathbb{P}(q_\alpha(X) \ne \bar{\lambda}_\alpha) > 0$, then by strict monotonicity of each F_t we also have

$$\mathbb{P}\big(F_T(\bar{\lambda}_\alpha) \ne 1 - \alpha\big) > 0.$$

If $C_t(\cdot)$ is strictly convex for almost every t, then the supporting-line inequality is strict on a set of positive probability, which implies

$$\mathbb{E}\big[G_X(\lambda^\star(X))\big] < \mathbb{E}\big[G_X(\bar{\lambda}_\alpha)\big].$$

□

B.4. Proof of Theorem 4.4

Proof. Let $T = \psi(X)$ and set

$$u_0 := 1 - \alpha, \qquad \lambda^\star(X) = q_\alpha(T) = F_T^{-1}(u_0).$$

By Proposition 4.3,

$$\mathbb{E}\big[G_X(\lambda^\star(X))\big] \le \mathbb{E}\big[G_X(\bar{\lambda}_\alpha)\big], \tag{B.9}$$

with strict inequality under the additional strictness assumptions stated there.

It remains to show that

$$\mathbb{E}\big[G_X(\widehat{\lambda}_{\alpha,N}(X))\big] \longrightarrow \mathbb{E}\big[G_X(\lambda^\star(X))\big].$$

Fix t. Since F_t is continuous and strictly increasing on $[0,1]$, we have

$$F_t^{-1}(F_t(\lambda)) = \lambda \qquad \text{for all } \lambda \in [0,1].$$

By definition,

$$C_t(u) = G_t(F_t^{-1}(u)),$$

so for every $\lambda \in [0,1]$,

$$G_t(\lambda) = C_t(F_t(\lambda)).$$

Now $u \mapsto C_t(u)$ is convex and differentiable on $(0,1)$ by Proposition 4.3, hence continuous on $(0,1)$. Since

$$F_t(\lambda^\star(t)) = F_t\big(F_t^{-1}(u_0)\big) = u_0 = 1-\alpha \in (0,1),$$

it follows that $\lambda \mapsto G_t(\lambda)$ is continuous at $\lambda^\star(t)$.

Now fix $x \in \mathcal{X}$. By the assumed uniform consistency,

$$\big|\widehat{\lambda}_{\alpha,N}(x) - \lambda^\star(x)\big| \le \sup_{z\in\mathcal{X}} \big|\widehat{\lambda}_{\alpha,N}(z) - \lambda^\star(z)\big| \xrightarrow{p} 0,$$

so

$$\widehat{\lambda}_{\alpha,N}(x) \xrightarrow{p} \lambda^\star(x).$$

Since $G_x(\cdot)$ is continuous at $\lambda^\star(x)$, the continuous mapping theorem yields

$$G_x\big(\widehat{\lambda}_{\alpha,N}(x)\big) \xrightarrow{p} G_x\big(\lambda^\star(x)\big).$$

For $\varepsilon > 0$, define

$$p_{N,\varepsilon}(x) := \mathbb{P}_{\mathcal{D}_{\text{cal}}}\Big(\Big|G_x\big(\widehat{\lambda}_{\alpha,N}(x)\big) - G_x\big(\lambda^\star(x)\big)\Big| > \varepsilon\Big).$$

Then $p_{N,\varepsilon}(x) \to 0$ for every fixed x, and $0 \le p_{N,\varepsilon}(x) \le 1$. Since the fresh test point X is independent of the calibration sample,

$$\mathbb{P}\Big(\Big|G_X\big(\widehat{\lambda}_{\alpha,N}(X)\big) - G_X\big(\lambda^\star(X)\big)\Big| > \varepsilon\Big) = \mathbb{E}_X\big[p_{N,\varepsilon}(X)\big] \longrightarrow 0$$

by dominated convergence. Therefore,

$$G_X\big(\widehat{\lambda}_{\alpha,N}(X)\big) \xrightarrow{p} G_X\big(\lambda^\star(X)\big).$$

Moreover,

$$0 \le G_X\big(\widehat{\lambda}_{\alpha,N}(X)\big) \le 1, \qquad 0 \le G_X\big(\lambda^\star(X)\big) \le 1,$$

so the sequence is uniformly integrable. Hence convergence in probability upgrades to convergence in L^1, and therefore

$$\lim_{N\to\infty} \mathbb{E}\big[G_X(\widehat{\lambda}_{\alpha,N}(X))\big] = \mathbb{E}\big[G_X(\lambda^\star(X))\big].$$

Combining this with (B.9) proves

$$\lim_{N\to\infty} \mathbb{E}\big[G_X(\widehat{\lambda}_{\alpha,N}(X))\big] \le \mathbb{E}\big[G_X(\bar{\lambda}_\alpha)\big],$$

with strict inequality under the additional strictness assumptions from Proposition 4.3.

Finally, conditional on X and a threshold λ, each of the M sampled candidates is accepted with probability $G_X(\lambda)$, so

$$\mathbb{E}\big[\,|\widehat{C}_{\alpha,N}(X)| \,\big|\, X, \widehat{\lambda}_{\alpha,N}(X)\big] = M\, G_X\big(\widehat{\lambda}_{\alpha,N}(X)\big).$$

Taking expectations gives

$$\mathbb{E}\big[\,|\widehat{C}_{\alpha,N}(X)|\,\big] = M\,\mathbb{E}\big[G_X(\widehat{\lambda}_{\alpha,N}(X))\big].$$

Passing to the limit yields

$$\lim_{N\to\infty} \mathbb{E}\big[|\widehat{C}_{\alpha,N}(X)|\big] = M\,\mathbb{E}\big[G_X(\lambda^\star(X))\big] \le M\,\mathbb{E}\big[G_X(\bar{\lambda}_\alpha)\big],$$

with strict inequality under the same additional assumptions.

□

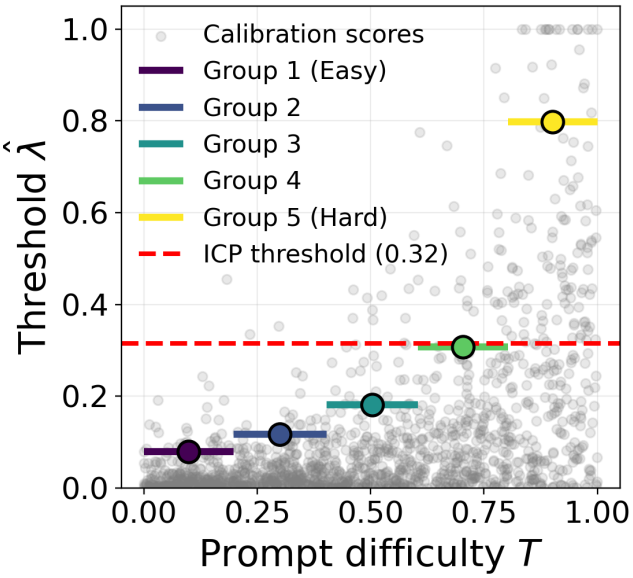

Figure 5. Learned threshold $\widehat{\lambda}_\alpha(X)$ versus prompt difficulty in the updated synthetic run ($\alpha = 0.10$, 5 bins). Easy prompts receive stricter thresholds, while harder prompts receive looser thresholds, explaining the improved group-wise coverage of CFC relative to global-threshold baselines.

C. Additional experiments

C.1. Synthetic Data

Setup details. All synthetic appendix results use the same clean synthetic generator as the main text. Each prompt has a scalar difficulty variable $T \in [0,1]$, we draw M candidates, and define the prompt-level latent success score as $S(X) = \min\{V_j : A_j = 1\}$, with $S(X) = 1$ if no sampled candidate is correct. Unless varied explicitly in the ablations, we use $N_{\text{cal}} = N_{\text{test}} = 10{,}000$ and $M = 50$. Throughout the synthetic appendix, ECR and GSC use the ground-truth correctness labels rather than the surrogate event $S(X) \le \widehat{\lambda}(X)$. The main synthetic comparisons use 10 equal-frequency bins for GSC, the threshold-adaptation plot below uses 5 bins for readability, and CFC-PAC uses the same stability-mode adjustment with $\delta = 0.90$ as in the main synthetic experiments.

Threshold adaptation versus prompt difficulty. Figure 5 visualizes the learned threshold as a function of prompt difficulty in the updated synthetic run at $\alpha = 0.10$ using 5 difficulty bins. As expected, CFC assigns stricter thresholds to easy prompts and looser thresholds to hard prompts. This is exactly the adaptive behavior that a single global-threshold baseline cannot express.

Full target-risk sweep. Table 7 reports the full synthetic sweep across target error rates. As above, ECR and GSC use ground-truth correctness labels, and CFC-PAC uses the stability-mode adjustment with $\delta = 0.90$ throughout. We include LEARNT CP in the full sweep to separate gains from learning a better threshold from gains due to exact conditional conformalization.

Sensitivity to calibration size and sampling budget. Table 8 reports a representative synthetic ablation of CFC and CFC-PAC as we vary the number of calibration points N_{cal} and the sampling budget M, using the same 10-bin setting as the main synthetic comparison. For consistency with the main synthetic experiments, CFC-PAC uses the same stability-mode adjustment with $\delta = 0.90$, and all reported coverages are true label-based coverages. Figure 6 shows the corresponding group-level miscoverage profile.

Table 7. Results at different target error rates α.

Methods	$\alpha = 0.10$			$\alpha = 0.15$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	90.6 ± 0.1	58.2 ± 1.3	16.00 ± 0.00	85.0 ± 1.6	44.9 ± 2.3	10.80 ± 0.98
ICP	90.2 ± 0.2	57.4 ± 1.4	16.71 ± 0.23	85.3 ± 0.6	46.5 ± 1.4	12.11 ± 0.28
Learnt CP	90.2 ± 0.3	84.3 ± 0.5	15.72 ± 0.15	85.2 ± 0.6	77.4 ± 0.6	12.44 ± 0.17
CFC (ours)	90.3 ± 0.5	88.7 ± 0.7	15.53 ± 0.12	85.2 ± 0.6	82.7 ± 0.8	12.42 ± 0.09
CFC-PAC (ours)	90.8 ± 0.6	89.1 ± 0.5	15.87 ± 0.16	85.6 ± 0.6	83.4 ± 1.1	12.66 ± 0.07
Methods	$\alpha = 0.20$			$\alpha = 0.25$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	79.7 ± 0.2	36.5 ± 0.9	8.00 ± 0.00	73.6 ± 0.2	29.4 ± 1.2	6.00 ± 0.00
ICP	80.3 ± 0.7	37.9 ± 1.6	9.32 ± 0.26	75.4 ± 0.8	31.6 ± 1.4	7.43 ± 0.21
Learnt CP	80.2 ± 0.6	71.2 ± 0.9	10.31 ± 0.09	75.4 ± 0.8	65.1 ± 1.6	8.72 ± 0.13
CFC (ours)	80.2 ± 0.6	77.4 ± 0.7	10.39 ± 0.07	75.3 ± 0.8	72.1 ± 1.2	8.80 ± 0.10
CFC-PAC (ours)	80.6 ± 0.6	78.1 ± 0.6	10.53 ± 0.07	75.7 ± 0.8	72.4 ± 1.2	8.89 ± 0.10
Methods	$\alpha = 0.30$			$\alpha = 0.35$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	73.6 ± 0.2	29.4 ± 1.2	6.00 ± 0.00	64.2 ± 0.3	21.7 ± 1.4	4.00 ± 0.00
ICP	70.5 ± 0.8	27.1 ± 1.5	6.00 ± 0.19	65.6 ± 0.7	23.1 ± 1.0	4.91 ± 0.16
Learnt CP	70.4 ± 0.9	59.6 ± 1.7	7.37 ± 0.12	65.5 ± 1.0	54.7 ± 1.6	6.32 ± 0.14
CFC (ours)	70.4 ± 0.9	66.5 ± 1.4	7.53 ± 0.11	65.5 ± 0.8	61.3 ± 1.8	6.48 ± 0.10
CFC-PAC (ours)	70.7 ± 0.9	66.8 ± 1.3	7.60 ± 0.11	65.8 ± 0.9	61.5 ± 1.9	6.55 ± 0.11
Methods	$\alpha = 0.40$			$\alpha = 0.45$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	64.2 ± 0.3	21.7 ± 1.4	4.00 ± 0.00	64.2 ± 0.3	21.7 ± 1.4	4.00 ± 0.00
ICP	60.8 ± 0.8	19.4 ± 0.7	4.06 ± 0.13	55.6 ± 1.0	16.5 ± 0.9	3.32 ± 0.13
Learnt CP	60.4 ± 0.9	49.6 ± 1.8	5.40 ± 0.13	55.7 ± 1.0	45.5 ± 1.6	4.66 ± 0.10
CFC (ours)	60.5 ± 0.7	56.4 ± 1.7	5.58 ± 0.08	55.6 ± 0.8	51.9 ± 1.5	4.83 ± 0.10
CFC-PAC (ours)	60.9 ± 0.7	56.6 ± 1.6	5.62 ± 0.09	56.0 ± 0.9	52.3 ± 1.5	4.88 ± 0.11

Table 8. Ablation of CFC vs. CFC-PAC on synthetic data, with 10 bins.

		CFC		CFC-PAC	
N_{cal}	M	True cov.	Mean set	True cov.	Mean set
2000	50	0.799 ± 0.006	10.60 ± 0.35	0.808 ± 0.005	10.92 ± 0.30
2000	100	0.809 ± 0.012	9.77 ± 0.31	0.819 ± 0.012	10.15 ± 0.29
2000	150	0.796 ± 0.006	9.03 ± 0.43	0.807 ± 0.006	9.37 ± 0.43
5000	50	0.801 ± 0.004	10.45 ± 0.12	0.808 ± 0.004	10.69 ± 0.16
5000	100	0.803 ± 0.008	9.52 ± 0.19	0.809 ± 0.007	9.74 ± 0.15
5000	150	0.802 ± 0.006	9.13 ± 0.26	0.808 ± 0.007	9.35 ± 0.27
10000	50	0.802 ± 0.006	10.39 ± 0.07	0.806 ± 0.006	10.53 ± 0.07
10000	100	0.803 ± 0.005	9.54 ± 0.06	0.808 ± 0.006	9.69 ± 0.09
10000	150	0.799 ± 0.008	9.15 ± 0.15	0.803 ± 0.008	9.29 ± 0.15

C.2. Real-World Data

C.2.1. TriviaQA

Full target-risk sweep. For the chosen TriviaQA feature map, we compute the rank-normalized answer-distribution entropy $T_{\text{ent}}(X)$ and the rank-normalized maximum verifier loss $T_{\text{loss}}(X)$ on the calibration split, then assign a prompt to the hard group when $\max\{T_{\text{ent}}(X), T_{\text{loss}}(X)\} \geq q_{0.925}$, where $q_{0.925}$ is the calibration 92.5th percentile of that combined score. Table 9 reports the full TriviaQA sweep under this chosen feature map. The full table makes the main-paper tradeoff more explicit: **CFC** is the most size-efficient of our methods, while **CFC-PAC-FULL** is the strongest at hitting the target coverage with a higher subgroup floor.

C.2.2. GSM8K

Full target-risk sweep. Table 10 reports the full GSM8K target-risk sweep for the chosen setting from the main paper: we keep the first 5 sampled candidates per prompt, define $T(X)$ as the mean verifier loss across those candidates, and use the quadratic basis $\Phi(X) = [1, T(X), T(X)^2]$. The same qualitative pattern holds across the sweep: the conditional methods

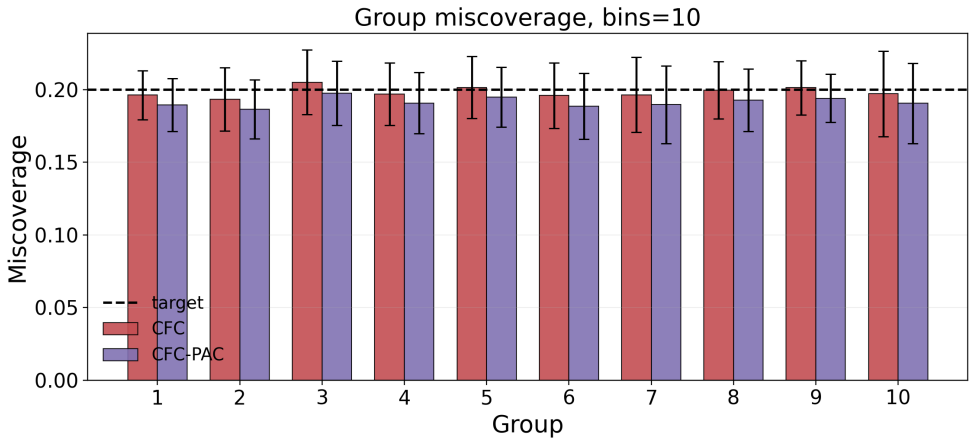

Figure 6. Group miscoverage for CFC and CFC-PAC, with 10 bins.

Table 9. Full TriviaQA sweep across target error rates α under the calibration-defined split $\max\{T_{\text{ent}}(X), T_{\text{loss}}(X)\} \geq q_{0.925}$. APSS below 1 indicates that the method abstains and returns the empty set on some prompts. ECR and GSC are reported in percent; for ECR, values closest to the target coverage $1 - \alpha$ are preferred.

Methods	$\alpha = 0.20$			$\alpha = 0.25$			$\alpha = 0.30$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	81.6 ± 0.3	69.7 ± 1.5	1.75 ± 0.00	73.4 ± 0.3	55.9 ± 1.6	1.00 ± 0.00	73.4 ± 0.3	55.9 ± 1.6	1.00 ± 0.00
ICP	80.0 ± 0.5	65.1 ± 1.5	1.43 ± 0.02	74.9 ± 0.3	56.7 ± 1.9	1.08 ± 0.01	69.5 ± 0.8	49.3 ± 2.7	0.90 ± 0.02
Learnt CP	79.6 ± 0.5	76.3 ± 1.6	1.69 ± 0.04	74.7 ± 0.4	74.0 ± 1.1	1.22 ± 0.03	69.5 ± 0.7	68.7 ± 1.3	0.97 ± 0.02
CFC (ours)	76.2 ± 0.5	65.4 ± 1.7	1.21 ± 0.01	72.7 ± 0.4	65.2 ± 1.8	1.03 ± 0.03	68.4 ± 0.7	62.8 ± 2.0	0.88 ± 0.01
CFC-PAC (ours)	76.4 ± 0.5	65.4 ± 1.7	1.22 ± 0.01	73.1 ± 0.5	65.3 ± 1.8	1.05 ± 0.04	68.8 ± 0.7	62.9 ± 2.1	0.89 ± 0.02
CFC-FULL (ours)	79.6 ± 0.5	76.3 ± 1.6	1.69 ± 0.04	74.8 ± 0.3	74.1 ± 1.1	1.26 ± 0.09	69.6 ± 0.7	68.9 ± 1.1	0.97 ± 0.02
CFC-PAC-FULL (ours)	80.1 ± 0.5	76.3 ± 1.6	1.72 ± 0.04	75.3 ± 0.4	74.6 ± 1.0	1.32 ± 0.10	70.0 ± 0.8	69.2 ± 1.3	0.99 ± 0.03
Methods	$\alpha = 0.35$			$\alpha = 0.40$			$\alpha = 0.45$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	73.4 ± 0.3	55.9 ± 1.6	1.00 ± 0.00	73.4 ± 0.3	55.9 ± 1.6	1.00 ± 0.00	73.4 ± 0.3	55.9 ± 1.6	1.00 ± 0.00
ICP	64.5 ± 0.6	43.0 ± 2.1	0.78 ± 0.01	59.9 ± 0.9	36.1 ± 3.1	0.70 ± 0.01	54.8 ± 0.9	29.6 ± 2.3	0.62 ± 0.01
Learnt CP	64.6 ± 0.7	63.0 ± 2.1	0.82 ± 0.02	59.9 ± 0.8	58.2 ± 2.4	0.72 ± 0.01	55.2 ± 1.0	53.8 ± 1.7	0.65 ± 0.01
CFC (ours)	63.9 ± 0.7	59.2 ± 2.3	0.78 ± 0.01	59.5 ± 0.9	55.8 ± 2.7	0.70 ± 0.01	54.9 ± 1.0	52.3 ± 2.4	0.63 ± 0.01
CFC-PAC (ours)	64.3 ± 0.7	59.5 ± 2.4	0.78 ± 0.01	59.9 ± 0.8	56.4 ± 2.9	0.70 ± 0.01	55.4 ± 1.0	52.8 ± 2.6	0.64 ± 0.01
CFC-FULL (ours)	64.7 ± 0.7	63.2 ± 2.0	0.82 ± 0.02	59.9 ± 0.9	58.4 ± 2.2	0.72 ± 0.01	55.2 ± 1.0	53.8 ± 1.7	0.65 ± 0.01
CFC-PAC-FULL (ours)	65.1 ± 0.8	63.7 ± 2.2	0.83 ± 0.02	60.4 ± 0.8	59.0 ± 2.2	0.73 ± 0.01	55.7 ± 1.0	54.3 ± 1.9	0.65 ± 0.01

remain much more efficient than ICP while sharply improving worst-group coverage.

Candidate-budget ablation. Table 11 compares the chosen $N = 5$ budget against $N = 20$ at the representative target $\alpha = 0.10$, keeping the same mean-loss proxy and quadratic basis. The larger candidate budget barely changes target calibration, but it inflates APSS substantially for every threshold-based method. This is why the main paper uses the smaller budget on GSM8K: the extra samples add little new diversity but materially hurt efficiency.

C.2.3. Flickr8k

Full target-risk sweep. Table 12 reports the full Flickr8k sweep for the chosen clean setting from the main paper: we keep up to two cached candidates per image, define $T(X)$ as the mean verifier loss across those candidates, and use the quadratic basis $\Phi(X) = [1, T(X), T(X)^2]$. This benchmark is visibly easier than GSM8K or TriviaQA, so the most informative comparison is closeness to the target coverage together with subgroup reliability. At $\alpha = 0.03$ (target coverage 97%), **CFC-PAC-FULL** is the closest full-set variant to target while improving GSC over every baseline, whereas **CFC** collapses to almost one caption per image and is best viewed as the smallest-size extreme.

Table 10. Full GSM8K sweep across target error rates α using the first five sampled candidates per prompt, $T(X)$ equal to mean verifier loss, and the quadratic basis $\Phi(X) = [1, T(X), T(X)^2]$. ECR and GSC are reported in percent; for ECR, values closest to the target coverage $1 - \alpha$ are preferred.

Methods	$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.15$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00
ICP	95.09 ± 1.42	79.85 ± 6.53	4.73 ± 0.09	90.39 ± 1.44	56.36 ± 6.20	4.36 ± 0.08	83.91 ± 2.15	27.73 ± 8.14	3.97 ± 0.11
Learnt CP	94.91 ± 1.03	88.48 ± 3.05	4.01 ± 0.98	90.09 ± 1.38	86.06 ± 0.77	2.22 ± 0.07	84.70 ± 1.11	79.09 ± 1.56	1.92 ± 0.06
CFC (ours)	94.82 ± 0.97	88.48 ± 2.32	2.35 ± 0.43	90.18 ± 1.41	86.36 ± 1.07	1.49 ± 0.04	84.97 ± 1.12	79.55 ± 1.52	1.34 ± 0.04
CFC-FULL (ours)	95.03 ± 0.97	88.94 ± 2.74	4.08 ± 0.93	90.30 ± 1.40	86.36 ± 1.07	2.23 ± 0.08	85.09 ± 1.03	79.55 ± 1.52	1.96 ± 0.06
CFC-PAC (ours)	95.03 ± 1.33	88.18 ± 2.47	2.59 ± 0.29	91.79 ± 1.66	88.18 ± 1.41	1.55 ± 0.06	86.64 ± 0.95	81.67 ± 1.11	1.38 ± 0.03
CFC-PAC-FULL (ours)	95.24 ± 1.40	88.79 ± 3.01	4.59 ± 0.62	91.91 ± 1.68	88.48 ± 1.62	2.34 ± 0.11	86.76 ± 0.88	81.67 ± 1.11	2.05 ± 0.05
Methods	$\alpha = 0.20$			$\alpha = 0.25$			$\alpha = 0.30$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00
ICP	79.55 ± 2.91	18.48 ± 4.92	3.69 ± 0.16	74.64 ± 2.53	13.48 ± 1.30	3.38 ± 0.15	70.18 ± 2.51	10.15 ± 0.91	3.11 ± 0.13
Learnt CP	79.42 ± 1.87	71.36 ± 3.63	1.74 ± 0.10	74.82 ± 2.51	66.97 ± 2.90	1.57 ± 0.07	69.36 ± 2.48	60.91 ± 2.23	1.42 ± 0.08
CFC (ours)	79.94 ± 1.80	72.12 ± 3.85	1.22 ± 0.04	75.03 ± 2.52	67.12 ± 2.86	1.12 ± 0.05	69.55 ± 2.53	61.21 ± 2.11	1.01 ± 0.05
CFC-FULL (ours)	80.06 ± 1.78	72.12 ± 3.85	1.76 ± 0.08	75.15 ± 2.47	67.12 ± 2.86	1.60 ± 0.07	69.64 ± 2.48	61.21 ± 2.11	1.44 ± 0.09
CFC-PAC (ours)	81.36 ± 1.68	73.18 ± 4.27	1.25 ± 0.04	75.94 ± 2.47	68.03 ± 2.81	1.13 ± 0.05	70.82 ± 2.58	62.58 ± 2.01	1.04 ± 0.05
CFC-PAC-FULL (ours)	81.48 ± 1.65	73.18 ± 4.27	1.82 ± 0.08	76.06 ± 2.42	68.03 ± 2.81	1.63 ± 0.07	70.94 ± 2.52	62.58 ± 2.01	1.48 ± 0.08

Table 11. GSM8K sample-budget ablation at $\alpha = 0.10$ under the mean-loss quadratic rule. Each budget uses the same basis $\Phi(X) = [1, T(X), T(X)^2]$, differing only in the number of retained sampled candidates per prompt.

Methods	$N = 5$			$N = 20$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	96.42 ± 0.39	86.52 ± 1.11	1.00 ± 0.00	96.70 ± 0.36	88.48 ± 1.11	1.00 ± 0.00
ICP	90.39 ± 1.44	56.36 ± 6.20	4.36 ± 0.08	90.15 ± 1.37	55.76 ± 5.52	16.75 ± 0.35
Learnt CP	90.09 ± 1.38	86.06 ± 0.77	2.22 ± 0.07	90.15 ± 0.81	87.42 ± 1.41	7.24 ± 0.33
CFC (ours)	90.18 ± 1.41	86.36 ± 1.07	1.49 ± 0.04	90.30 ± 0.63	87.73 ± 1.47	3.79 ± 0.12
CFC-FULL (ours)	90.30 ± 1.40	86.36 ± 1.07	2.23 ± 0.08	90.45 ± 0.70	87.73 ± 1.47	7.50 ± 0.18
CFC-PAC (ours)	91.79 ± 1.66	88.18 ± 1.41	1.55 ± 0.06	91.91 ± 0.46	88.64 ± 2.30	3.99 ± 0.07
CFC-PAC-FULL (ours)	91.91 ± 1.68	88.48 ± 1.62	2.34 ± 0.11	92.06 ± 0.57	88.79 ± 2.46	7.97 ± 0.03

Table 12. Full Flickr8k sweep with QWEN2-VL-7B-INSTRUCT using up to two cached candidates per image, $T(X)$ equal to mean verifier loss, and the quadratic basis $\Phi(X) = [1, T(X), T(X)^2]$. ECR and GSC are reported in percent; for ECR, values closest to the target coverage $1 - \alpha$ are preferred.

Methods	$\alpha = 0.01$			$\alpha = 0.02$			$\alpha = 0.03$		
	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓	ECR	GSC↑	APSS↓
TopK	97.75 ± 0.19	95.62 ± 0.71	2.00 ± 0.00	97.75 ± 0.19	95.62 ± 0.71	2.00 ± 0.00	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00
ICP	97.29 ± 0.29	93.33 ± 1.45	1.93 ± 0.01	96.25 ± 0.66	88.23 ± 3.66	1.87 ± 0.02	95.58 ± 0.54	85.21 ± 3.14	1.84 ± 0.01
Learnt CP	97.39 ± 0.33	95.52 ± 0.85	1.74 ± 0.06	97.06 ± 0.31	95.10 ± 0.63	1.34 ± 0.07	96.16 ± 0.17	94.48 ± 0.26	1.22 ± 0.07
CFC (ours)	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00	95.81 ± 0.38	93.23 ± 0.47	0.99 ± 0.00
CFC-FULL (ours)	97.66 ± 0.16	95.62 ± 0.71	1.86 ± 0.04	97.27 ± 0.21	95.21 ± 0.77	1.40 ± 0.06	96.27 ± 0.24	94.58 ± 0.26	1.25 ± 0.08
CFC-PAC (ours)	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00	96.37 ± 0.17	93.23 ± 0.47	1.00 ± 0.00
CFC-PAC-FULL (ours)	97.75 ± 0.19	95.62 ± 0.71	2.00 ± 0.00	97.64 ± 0.13	95.62 ± 0.71	1.86 ± 0.06	97.27 ± 0.21	95.21 ± 0.77	1.42 ± 0.07

Table 13. Flickr8k setting ablation at $\alpha = 0.03$ (target coverage 97%). Each cell reports ECR/APSS/GSC for the corresponding method.

Setting	CFC	CFC-PAC-FULL
$N = 2$, max-loss, poly2	96.02 / 1.00 / 93.02	97.62 / 1.89 / 95.93
Chosen: $N = 2$, mean-loss, poly2	95.81 / 0.99 / 93.23	97.27 / 1.42 / 95.21
$N = 3$, mean-loss, poly2	96.00 / 1.00 / 93.02	97.81 / 2.16 / 96.35

Setting ablation. Table 13 compares the chosen Flickr8k setting against two nearby alternatives from the clean search at the representative target $\alpha = 0.03$. The chosen $N = 2$ mean-loss quadratic rule is the best-balanced option we found: it keeps the **CFC** variant essentially at single-caption size, while **CFC-PAC-FULL** remains close to target without the larger APSS jump of the $N = 3$ alternative.