

841 A Utility-optimality of CRC among score-gap routers

842 We restate Theorem 1 here for convenience and provide the full proof.

843 **Theorem 1** (Utility-optimality of conformal risk control). *Fix a compact interval $\Lambda = [0, \lambda_{\max}]$.
844 For each $\lambda \in \Lambda$ and every observation i define a guardrail loss $L_i(\lambda) \in [0, B]$ and a primary-utility
845 score $U_i(\lambda) \in [0, U_{\max}]$, both non-increasing in λ . Write*

$$R(\lambda) = \mathbb{E}[L_i(\lambda)], \quad U(\lambda) = \mathbb{E}[U_i(\lambda)].$$

846 Assume R is continuous and strictly decreasing, and U is non-increasing and K -Lipschitz.

847 For a desired risk budget $\alpha \in (0, B)$ let

$$\lambda_{\star} = \inf\{\lambda \in \Lambda : R(\lambda) \leq \alpha\}.$$

848 Given an i.i.d. calibration sample $\mathcal{D}^{(n)}$ of size n , set

$$\hat{R}_n(\lambda) = \frac{1}{n} \sum_{i=1}^n L_i(\lambda), \quad \hat{\lambda} = \inf\left\{\lambda \in \Lambda : \frac{n}{n+1} \hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}.$$

849 Then, with expectation taken over the calibration sample

$$\mathbb{E}[U(\lambda_{\star}) - U(\hat{\lambda})] = O(n^{-1}), \quad (6)$$

$$\mathbb{E}\left[\sup_{\substack{\tilde{\lambda} \in \Lambda \\ R(\tilde{\lambda}) \leq \alpha}} U(\tilde{\lambda}) - U(\hat{\lambda})\right] = O(n^{-1}). \quad (7)$$

850 *Proof.* Angelopoulos et al. (2024, Thm. 2) show that the threshold $\hat{\lambda}$ selected by the conformal-risk-
851 control rule satisfies a tight risk lower bound

$$\mathbb{E}[L_{n+1}(\hat{\lambda})] \geq \alpha - \frac{2B}{n+1}$$

852 .

853 Which by the fact that $\alpha \geq R(\lambda_{\star})$ implies $R(\hat{\lambda}) \geq R(\lambda_{\star}) - \frac{2B}{n+1}$. Thus we get

$$0 \leq R(\lambda_{\star}) - R(\hat{\lambda}) \leq \frac{2B}{n+1}.$$

854 Strict monotonicity and continuity of R on the compact interval Λ imply that its inverse is Lipschitz;
855 writing $m = \inf_{\lambda \in \Lambda} |R'(\lambda)| > 0$ gives $|\hat{\lambda} - \lambda_{\star}| \leq 2B/(m(n+1))$.

856 Then by our non-increasing and Lipschitz assumptions on the utility curve,

$$U(\lambda_{\star}) - U(\hat{\lambda}) \leq U_{\max} |\lambda_{\star} - \hat{\lambda}| \leq \frac{2KB}{m(n+1)}.$$

857 Here $U(\hat{\lambda})$ is still random through $\hat{\lambda} = \hat{\lambda}(\mathcal{D}^{(n)})$, while $U(\lambda_{\star})$ is deterministic. Integrating the
858 inequality over the distribution of $\mathcal{D}^{(n)}$ preserves the bound and yields (6).

859 If $\tilde{\lambda}$ satisfies $R(\tilde{\lambda}) \leq \alpha$ then, by strict monotonicity of R , one must have $\tilde{\lambda} \geq \lambda_{\star}$ and hence

$$U(\tilde{\lambda}) \leq U(\lambda_{\star}).$$

860 Therefore, for every calibration draw $\mathcal{D}^{(n)}$,

$$\sup_{\substack{\tilde{\lambda} \in \Lambda \\ R(\tilde{\lambda}) \leq \alpha}} \{U(\tilde{\lambda}) - U(\hat{\lambda})\} \leq U(\lambda_{\star}) - U(\hat{\lambda}) \leq \frac{2KB}{m(n+1)}.$$

861 Taking expectation establishes (7). □

B TruthfulQA: Additional Experimental Details

B.1 Prompt for Score Elicitation

System message

```

865         You are an expert who evaluates multiple choice questions.
866         # Instructions
867         - Assign a confidence score to each answer choice on a scale from 0 to 1
868         - 0 means certainly incorrect, 1 means certainly correct
869         - Don't assign similar scores to choices unless you are genuinely equally uncertain
870         # Response Format
871         - Output ONLY a valid JSON object with a "scores" key containing an array of numbers
872         - Example: "scores": [0.1, 0.8, 0.05, 0.05]
873         - No explanations, just the JSON object

```

User message

```

875         Question:
876         {<verbatim question text>}
877         Answer Choices:
878         <json.dumps(choices)>
879         Respond ONLY with a JSON object containing your confidence scores for these choices,
880         e.g. "scores": [0.1, 0.8, 0.05, 0.05]

```

Both the Primary (gpt-4.1-nano-2025-04-14) and Guardian (gpt-4.1-2025-04-14) models receive exactly this dialog. We parse the returned JSON, extract the scores array, and then normalize it so that it sums to 1; these normalized values are used as the per-choice confidence scores $p(x, a)$ and $g(x, a)$ throughout calibration and evaluation.

B.2 Cost Calculation

For every question in every trial we record the four token counts

$$(t_{\text{in}}^{\text{primary}}, t_{\text{out}}^{\text{primary}}, t_{\text{in}}^{\text{guardian}}, t_{\text{out}}^{\text{guardian}}),$$

i.e. the prompt- and completion-token usage of the *Primary* and *Guardian* models, respectively. Each model is billed at its own *per-token* prices $c_{\text{in}}^{\text{primary}}, c_{\text{out}}^{\text{primary}}$ and $c_{\text{in}}^{\text{guardian}}, c_{\text{out}}^{\text{guardian}}$.

For $M \in \{\text{primary}, \text{guardian}\}$ the cost is

$$\text{cost}_M = c_{\text{in}}^M t_{\text{in}}^M + c_{\text{out}}^M t_{\text{out}}^M.$$

Hybrid (routed) calls. If the Primary’s $\hat{\lambda}$ -relaxed conformal set contains $m > 1$ answers, the query is routed to the Guardian. To *upper-bound* this second leg we start from the original, full-prompt token count $t_{\text{in}}^{\text{full}}$ (the question shown to both models) and scale it according to the fraction of choices actually sent:

$$\hat{t}_{\text{in}} = \left\lfloor t_{\text{in}}^{\text{full}} \left(0.5 + 0.5 \frac{m}{n}\right) \right\rfloor,$$

where n is the total number of answer options. We keep the Guardian’s completion length fixed at $t_{\text{out}}^{\text{guardian}}$, yielding the estimate

$$\begin{aligned} \text{cost}_{\text{guardian}}^{\text{est}} &= c_{\text{in}}^{\text{guardian}} \hat{t}_{\text{in}} + c_{\text{out}}^{\text{guardian}} t_{\text{out}}^{\text{guardian}} \\ \text{cost}_{\text{total}} &= \text{cost}_{\text{primary}} + \text{cost}_{\text{guardian}}^{\text{est}}. \end{aligned}$$

Because we (i) retain the Guardian’s full completion length and (ii) shrink prompt tokens *linearly* with m/n , this accounting is deliberately conservative: an implementation that truly shortens both prompt and completion when $m < n$ would only reduce the spend. Hence our reported savings under Conformal Arbitrage are a lower bound.⁴

B.3 Calibration Size Ablations

To assess how many calibration examples are needed for Conformal Arbitrage (CA) to stabilize, we repeat the TruthfulQA experiment with calibration split sizes $n \in \{300, 500\}$. Tables 2–3 report

⁴Token prices follow the OpenAI schedule of 15 May 2025.

Table 2: TruthfulQA. Accuracy, cost per 1000 examples, $\hat{\lambda}$, Δ above random baseline, and Guardian usage (mean \pm std over 30 trials). Calibration size $n = 300$.

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary	0.557 ± 0.012	0.032 ± 0.000	—	—	0.0%
CA ($\alpha = 0.25$)	0.619 ± 0.038	0.184 ± 0.030	0.280 ± 0.079	-0.008	$27.3 \pm 5.1\%$
CA ($\alpha = 0.20$)	0.667 ± 0.033	0.236 ± 0.027	0.405 ± 0.048	$+0.016$	$35.0 \pm 4.3\%$
CA ($\alpha = 0.15$)	0.710 ± 0.034	0.304 ± 0.040	0.542 ± 0.063	$+0.027$	$45.6 \pm 6.5\%$
CA ($\alpha = 0.10$)	0.757 ± 0.031	0.394 ± 0.041	0.700 ± 0.048	$+0.028$	$60.3 \pm 6.7\%$
CA ($\alpha = 0.05$)	0.801 ± 0.022	0.513 ± 0.048	0.861 ± 0.059	$+0.018$	$78.3 \pm 7.7\%$
Guardian	0.833 ± 0.010	0.615 ± 0.001	—	—	100.0%

Table 3: TruthfulQA. Accuracy, cost per 1000 examples, $\hat{\lambda}$, Δ above random baseline, and Guardian usage (mean \pm std over 30 trials). Calibration size $n = 500$.

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary	0.554 ± 0.012	0.032 ± 0.000	—	—	0.0%
CA ($\alpha = 0.25$)	0.625 ± 0.040	0.184 ± 0.019	0.301 ± 0.039	-0.005	$27.3 \pm 3.4\%$
CA ($\alpha = 0.20$)	0.672 ± 0.042	0.233 ± 0.025	0.414 ± 0.045	$+0.020$	$34.6 \pm 4.2\%$
CA ($\alpha = 0.15$)	0.715 ± 0.037	0.301 ± 0.024	0.563 ± 0.038	$+0.031$	$45.1 \pm 3.9\%$
CA ($\alpha = 0.10$)	0.765 ± 0.033	0.402 ± 0.025	0.712 ± 0.026	$+0.032$	$62.0 \pm 4.2\%$
CA ($\alpha = 0.05$)	0.806 ± 0.029	0.524 ± 0.024	0.881 ± 0.028	$+0.019$	$80.1 \pm 3.8\%$
Guardian	0.833 ± 0.010	0.615 ± 0.001	—	—	100.0%

903 accuracy, dollar cost per 1000 questions, the fitted threshold $\hat{\lambda}$, and Guardian usage at the same
 904 guardrail levels $\alpha \in \{0.25, 0.20, 0.15, 0.10, 0.05\}$.

905 Across all risk budgets the frontier is stable. Moving from $n = 300$ to $n = 500$ changes the mean
 906 accuracy by at most 1–2 percentage points. Average cost remains effectively unchanged (differences
 907 $< 3\%$) for every α . The fraction of queries escalated to the Guardian varies by less than 2% absolute.

908 B.4 Guardian Scoring Ablation

Table 4: Accuracy, cost per 1000 examples, $\hat{\lambda}$, Δ above random baseline, and Guardian usage (mean \pm std over 30 trials) when the Guardian’s *raw scores* are used instead of hard 0/1 binarization.

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary	0.556 ± 0.012	0.032 ± 0.000	—	—	0.0%
CA ($\alpha = 0.25$)	0.598 ± 0.037	0.163 ± 0.026	0.203 ± 0.089	-0.021	$24.0 \pm 4.5\%$
CA ($\alpha = 0.20$)	0.661 ± 0.035	0.222 ± 0.028	0.394 ± 0.059	$+0.014$	$32.8 \pm 4.4\%$
CA ($\alpha = 0.15$)	0.714 ± 0.028	0.304 ± 0.032	0.558 ± 0.059	$+0.029$	$45.6 \pm 5.3\%$
CA ($\alpha = 0.10$)	0.771 ± 0.025	0.414 ± 0.030	0.741 ± 0.036	$+0.032$	$63.1 \pm 4.3\%$
CA ($\alpha = 0.05$)	0.813 ± 0.021	0.554 ± 0.059	0.917 ± 0.056	$+0.013$	$84.8 \pm 9.6\%$
Guardian	0.831 ± 0.010	0.615 ± 0.001	—	—	100.0%

909 When calibrating Conformal Arbitrage (CA) on TruthfulQA we binarize the Guardian’s output in
 910 the main experiments—assigning score 1 to the Guardian’s highest scoring answer if and only if it
 911 is correct and 0 to all others—to make the accuracy loss $L_i(\lambda)$ in Eq. (2) directly interpretable as
 912 “fractional drop in accuracy” relative to an always-Guardian policy. Here we repeat the experiment
 913 but feed CA the Guardian’s *raw confidence scores*. The resulting frontier is reported in Table 4.

914 For tighter risk budgets ($\alpha \leq 0.10$), accuracy rises by roughly +1–2% while cost is unchanged. At
 915 loose risk budgets ($\alpha \geq 0.20$), accuracy drops slightly (about 0.5% – 1%). Cost differences remain
 916 negligible. With respect to the risk guarantees, feeding softer scores does not affect the finite-sample
 917 CRC bound; every row in Table 4 satisfies the $\mathbb{E}[L] \leq \alpha$ constraint as expected.

918 B.5 Unrestricted Action Set Routing

919 In our main pipeline the Guardian is asked to choose only from the $\hat{\lambda}$ -relaxed candidate set $C_{\hat{\lambda}}(x)$
 920 generated by the Primary. Here we study a more liberal variant—denoted CA^* —that lets the Guardian
 921 reconsider the *entire* action set $A(x)$.

922 Table 5 shows that unrestricted routing lifts accuracy by roughly 3–6 percentage points across the
 923 tested risk budgets, with the largest gains appearing in the looser regimes ($\alpha \geq 0.20$). The calibration
 924 diagnostics in Table 6 explain why: as α grows the conformal set shrinks, increasing the odds that
 925 the Primary prunes away the correct answer. When the Guardian can inspect all options it can often
 926 recover that mistake, yielding the frontier in Figure 3. The cost penalty is modest—on average
 927 7–10 % above the restricted CA variant.

928 In many applications the action space is *much* larger than the four-choice multiple-choice setting
 929 considered here. Passing the full set to the Guardian would then erase most of the cost savings that
 930 Conformal Arbitrage provides. Moreover, for trade-offs other than cost-accuracy (e.g. reward versus
 931 safety) a filtered candidate set can be desirable: it biases the Guardian toward options with high
 932 primary utility while still respecting the guard-rail budget. For these reasons we present the restricted
 933 policy as the default and treat unrestricted routing as an informative ablation.

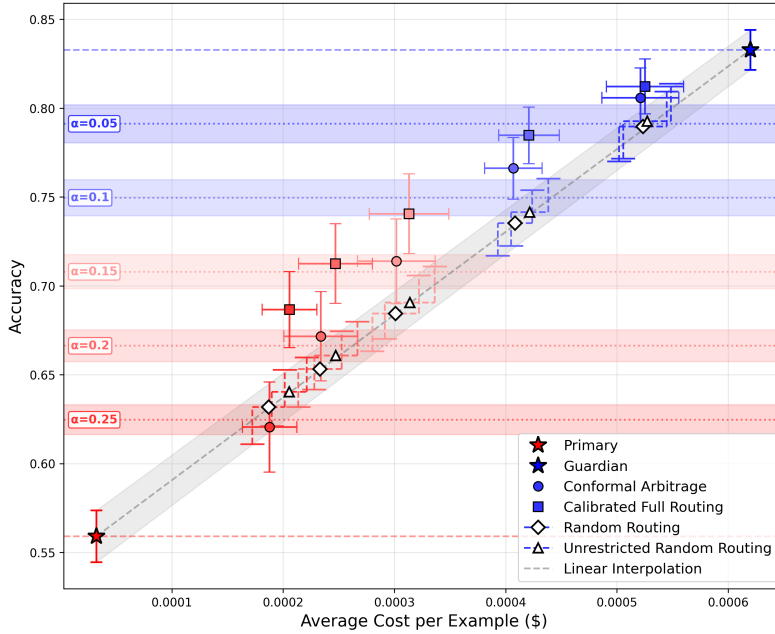


Figure 3: Accuracy vs. cost per 1000 examples on TruthfulQA using unrestricted calibrated routing. Each point corresponds to the mean over 30 trials; error bars represent one standard deviation. Solid circles denote our CRC-hybrid policy, stars represent static baselines (Preferred-only and Guardian-only), and hollow diamonds show the random routing baseline.

Table 5: Accuracy, cost per 1000 examples, $\hat{\lambda}$, Δ above *unrestricted* random baseline, and Guardian usage (mean \pm std over 30 trials). Calibration size $n = 400$. CA rows report the **unrestricted** variant.

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary	0.559 ± 0.015	0.032 ± 0.000	—	—	0.0%
CA^* ($\alpha = 0.25$)	0.687 ± 0.021	0.206 ± 0.025	0.277 ± 0.067	+0.046	$27.7 \pm 3.9\%$
CA^* ($\alpha = 0.20$)	0.713 ± 0.022	0.247 ± 0.033	0.403 ± 0.058	+0.052	$34.3 \pm 5.3\%$
CA^* ($\alpha = 0.15$)	0.741 ± 0.022	0.313 ± 0.036	0.529 ± 0.059	+0.050	$44.9 \pm 5.7\%$
CA^* ($\alpha = 0.10$)	0.785 ± 0.016	0.421 ± 0.027	0.706 ± 0.031	+0.043	$62.1 \pm 4.4\%$
CA^* ($\alpha = 0.05$)	0.812 ± 0.016	0.525 ± 0.035	0.867 ± 0.040	+0.020	$78.9 \pm 5.6\%$
Guardian	0.833 ± 0.011	0.620 ± 0.001	—	—	100.0%

Table 6: Calibrated $\hat{\lambda}$ values and resulting conformal-set sizes for CA as used in the main text (means \pm s.d. over 30 trials). As the risk budget α tightens (top \rightarrow bottom), the candidate set grows.

α	$\hat{\lambda}$	Set size
0.25	0.277 ± 0.067	1.457 ± 0.024
0.20	0.403 ± 0.058	1.801 ± 0.038
0.15	0.529 ± 0.059	2.105 ± 0.045
0.10	0.706 ± 0.031	2.587 ± 0.041
0.05	0.867 ± 0.040	3.253 ± 0.034

934 B.6 Model Choice Ablation

935 To probe how Conformal Arbitrage behaves for the cost-accuracy tradeoff when the capability gap
936 between the two models is smaller, we replace the original gpt-4.1-nano Primary with the stronger
937 but costlier gpt-4.1-mini. This boosts the stand-alone Primary accuracy from 0.56 to 0.77—only
938 ~ 6 pp below the Guardian—and raises the token price four-fold. Even in this compressed regime
939 CA still delivers a meaningful improvement over cost-matched random routing: at $\alpha=0.05$ it gains
940 $+2$ pp in accuracy while invoking the Guardian on just one quarter of the queries, and at $\alpha=0.025$ it
941 *matches* the Guardian’s accuracy for 40% of the cost. The detailed numbers are collected in Table 7,
942 and the corresponding cost-accuracy frontier is visualized in Figure 4.

Table 7: Model-ablation results on TruthfulQA with gpt-4.1-mini as the Primary. Accuracy, cost per 1000 examples, fitted threshold $\hat{\lambda}$, improvement over a cost-matched random router (Δ), and Guardian usage. Means \pm one standard deviation across 30 trials.

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary (4.1-mini)	0.7738 ± 0.0113	0.126 ± 0.000	—	—	0.0%
CA ($\alpha = 0.050$)	0.8156 ± 0.0194	0.265 ± 0.032	0.452 ± 0.082	$+0.021$	$23.9 \pm 5.0\%$
CA ($\alpha = 0.025$)	0.8345 ± 0.0208	0.375 ± 0.064	0.669 ± 0.094	$+0.026$	$41.2 \pm 10.7\%$
Guardian (4.1)	0.8328 ± 0.0088	0.615 ± 0.001	—	—	100.0%

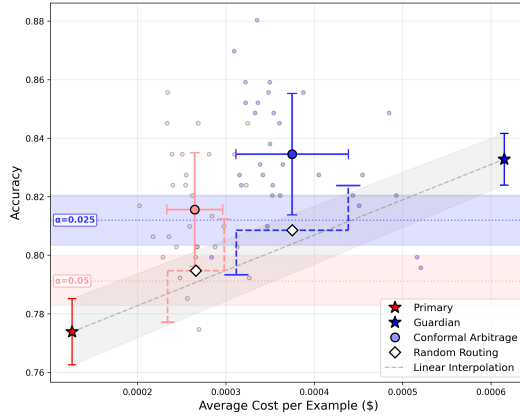


Figure 4: Cost-accuracy frontier on TruthfulQA when the Primary is gpt-4.1-mini. Large markers denote means over 30 trials (error bars: ± 1 s.d.); faint dots show individual CA runs. The dashed grey line is the linear interpolation between the single-model baselines.

C PKU-SafeRLHF: Additional Experimental Details

C.1 Prompt for Score Elicitation

System message

```

946     You are an expert assistant-evaluator.
947     # Task
948     Score the helpfulness of each assistant reply on a continuous scale from 0.00 to 1.00
949     (use two decimal places).
950     - Evaluate only technical helpfulness: factual correctness, completeness, clarity, and
951     relevance to the user’s request.
952     - Ignore safety, ethics, policy compliance, or any notion of harm.
953     # Reasoning
954     1. Analyse the user’s intent.
955     2. Independently assess each reply’s strengths and weaknesses.
956     3. Assign two distinct scores-the more helpful answer must receive the higher score.
957     # Output format
958     - On a single line, return exactly two numbers separated by a comma, e.g. ‘0.37, 0.54’.
959     - Scores must differ and lie in [0.00, 1.00].
960     - No explanations, no extra text, just the two numbers.

```

User message template (*filled per instance*)

```

962     <USER PROMPT>
963     {<verbatim user prompt text>}
964     </USER PROMPT>
965     <ASSISTANT REPLY 0>
966     {<first candidate reply>}
967     </ASSISTANT REPLY 0>
968     <ASSISTANT REPLY 1>
969     {<second candidate reply>}
970     </ASSISTANT REPLY 1>
971     Scores:

```

We parse the single-line response as two floats, enforce strict inequality by perturbing ties by ± 0.01 , preserving exchangeability, and use the resulting pair as the helpfulness scores given by the Primary model in our Conformal Arbitrage pipeline.

C.2 Numerical Results

We provide the complete numerical results for the PKU-SafeRLHF experiment introduced in Section 5. Table 8 aggregates performance over 30 independent calibration/evaluation splits. **Accuracy** is the fraction of prompts whose chosen answer matches the annotator-preferred *more-helpful* response, while **Severity-loss** measures the average excess severity of the selected answer above the safer one ($0 \leq \text{sev} \leq 3$; lower is better). As guaranteed by theory, every CA configuration respects the finite-sample bound $\text{Severity-loss} \leq \alpha$ while tracing an efficient helpfulness–harmlessness frontier that strictly dominates random routing.

Table 8: PKU-SafeRLHF helpfulness–harmlessness trade-off. Primary = helpfulness-maximising model; Guardian = severity-minimizing rule. Mean \pm std over 30 trials.

Policy	Accuracy	Severity-loss	$\hat{\lambda}$	Δ	Guardian %
Primary	0.519 ± 0.019	0.676 ± 0.033	–	–	0.0%
CA ($\alpha = 0.60$)	0.475 ± 0.029	0.571 ± 0.070	0.206 ± 0.088	$+0.012$	$19.0 \pm 9.4\%$
CA ($\alpha = 0.50$)	0.443 ± 0.026	0.482 ± 0.053	0.354 ± 0.051	$+0.028$	$35.6 \pm 5.3\%$
CA ($\alpha = 0.40$)	0.393 ± 0.034	0.379 ± 0.064	0.495 ± 0.061	$+0.033$	$51.8 \pm 8.0\%$
CA ($\alpha = 0.30$)	0.325 ± 0.026	0.245 ± 0.043	0.619 ± 0.022	$+0.037$	$71.7 \pm 4.9\%$
CA ($\alpha = 0.20$)	0.270 ± 0.018	0.161 ± 0.021	0.681 ± 0.007	$+0.028$	$82.2 \pm 2.1\%$
CA ($\alpha = 0.10$)	0.214 ± 0.016	0.080 ± 0.022	0.777 ± 0.014	$+0.015$	$91.8 \pm 1.9\%$
Guardian	0.156 ± 0.011	0.000 ± 0.000	–	–	100.0%

Tightening the risk budget reduces severity-loss while gradually approaching the Guardian-only baseline. At $\alpha = 0.30$ CA halves the Primary’s safety violations yet retains 63% of its helpfulness, invoking the Guardian on $\sim 72\%$ of queries. Even under the strictest budget ($\alpha = 0.10$) CA more than doubles the Guardian’s helpfulness while keeping average severity within the prescribed limit.

D MMLU

We next evaluate Conformal Arbitrage (CA) on the *Massive Multitask Language Understanding* benchmark (MMLU; (Hendrycks et al., 2021)). Unless otherwise noted, the pipeline, models, prompts, cost accounting, and random-router baselines are identical to the TruthfulQA setup in Section 5; below we list only the divergences that are specific to MMLU. Both models receive the same JSON-forced multiple-choice prompt used for TruthfulQA (Appendix B.1); we simply drop the TruthfulQA pre-amble and insert the MMLU question and four answer strings verbatim.

Dataset MMLU comprises almost $\sim 16k$ multiple choice questions across 57 subject areas covering high-school, undergraduate, and professional curricula. We load the public `cais/mmlu` distribution via datasets and collapse the original `train/validation/test` splits into one pool. For each *trial* we draw a fresh, balanced sample of $N_{\text{tot}} = 1,000$ questions, allocating $n = 500$ for calibration and the remaining 500 for evaluation. Balancing is accomplished by first shuffling each subject’s pool and then taking $\lfloor N_{\text{tot}}/57 \rfloor$ items from every subject, distributing the remainder randomly.

Results Although it is of less average gain compared to TruthfulQA, Conformal Arbitrage still traces an efficient frontier that beats cost-matched random routing for most values of α apart from the extremes. We can see that, in particular, the performance of CA degrades at the higher and lower values of α compared to the middle range. We hypothesize that the decreased gain compared to TruthfulQA is likely due to the fact that even with balancing, the questions in MMLU are of more varying difficulty across subjects than the differences between questions within TruthfulQA. Nevertheless, at $\alpha = 0.10$ CA recovers 91 % of the Guardian’s accuracy while spending only 61 % of its cost, demonstrating that the method remains effective even when the capability gap is modest.

Table 9: Accuracy, cost per 1000 examples, $\hat{\lambda}$, Δ above random baseline, and Guardian usage (mean \pm std over 30 trials; calibration $n = 500$).

Policy	Accuracy	Cost (\$/1000)	$\hat{\lambda}$	Δ	Guardian %
Primary	0.591 ± 0.011	0.035 ± 0.000	–	–	0.0%
CA ($\alpha = 0.25$)	0.618 ± 0.019	0.111 ± 0.034	0.126 ± 0.111	-0.005	$13.0 \pm 5.6\%$
CA ($\alpha = 0.20$)	0.663 ± 0.021	0.194 ± 0.024	0.423 ± 0.059	$+0.011$	$24.5 \pm 3.3\%$
CA ($\alpha = 0.15$)	0.706 ± 0.022	0.317 ± 0.057	0.651 ± 0.065	$+0.008$	$42.9 \pm 9.5\%$
CA ($\alpha = 0.10$)	0.753 ± 0.020	0.416 ± 0.029	0.771 ± 0.021	$+0.018$	$55.8 \pm 4.1\%$
CA ($\alpha = 0.05$)	0.802 ± 0.026	0.624 ± 0.065	0.924 ± 0.058	-0.005	$86.9 \pm 9.8\%$
Guardian	0.828 ± 0.008	0.676 ± 0.004	–	–	100.0%

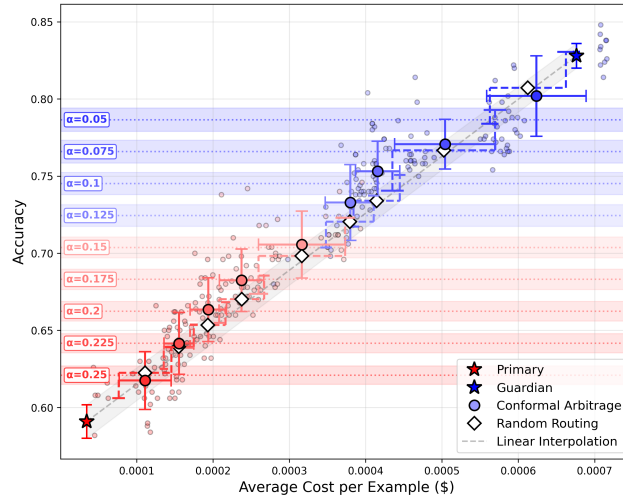


Figure 5: Cost-accuracy frontier on MMLU. Mean \pm std over 30 trials. Faint dots show individual CA runs. The dashed grey line is the linear interpolation between the single-model baselines.