Inference-Aware Prompt Optimization for Aligning Black-Box Large Language Models

Anonymous submission

Appendix A

Proof of Theorem 1

Theorem 1 (Error of PSST)**.** *Let $R = \lceil \log_2 |\mathcal{A}| \rceil$ be the number of trimming rounds, assume $[o_k^{\min}, o_k^{\max}] = [-1, 1]$, and define the* cost–gap complexity

$$H_1^c = \max_{(c,a^{c,i}) \neq (c,a^{c,1})} \frac{\bar{N}_{\max}}{\Delta^2_{c,a^{c,i}}}, \quad H_1 = \max_c H_1^c,$$

$$H_2^c = \max_{(c,a^{c,i}) \neq (c,a^{c,1})} \frac{i\bar{N}_{\max}}{\Delta^2_{c,a^{c,i}}}, \quad H_2 = \max_c H_2^c,$$

where $\Delta_{c,a^{c,i}} = Q^\alpha(c, a^{c,1}) - Q^\alpha(c, a^{c,i})$. Under a context c, arms are indexed based on ascending order of $Q^\alpha(c, a)$ and $\bar{N}_{\max}$ is defined as $\frac{N(a^{c,1})+N_{\max}}{2}$. Here, $N(a^{c,i})$ is the number of completions generated by the i-th indexed arm. Running PSST *with the structure-aware allocation for a total completion budget T returns the optimal arm in* every *context with probability at least*

$$1 - 3|\mathcal{C}|R \exp\left(-\frac{T}{\min\big(2|\mathcal{P}|H_1,\ 8|\mathcal{C}|H_2\big)\, R}\right).$$

Equivalently, to ensure failure probability at most δ it suffices to choose

$$T = O\Big(\min\big(|\mathcal{P}|H_1,\ |\mathcal{C}|H_2\big)\, R\, \log\Big(\tfrac{|\mathcal{C}|R}{\delta}\Big)\Big).$$

Indexing convention. Although Theorem 1 states that arms are indexed in ascending order of $Q^\alpha(c, a)$, throughout the proofs below we adopt the equivalent convention that $a^{c,1}$ denotes an optimal arm for context c, i.e., $Q^\alpha(c, a^{c,1}) = \max_a Q^\alpha(c, a)$, and thus $Q^\alpha(c, a^{c,1}) \geq Q^\alpha(c, a^{c,2}) \geq \cdots$. All quantities (e.g., $\Delta_{c,a^{c,i}}$ and $\bar{N}_i$) are interpreted under this convention. In particular, $\Delta_{c,a^{c,i}} \geq 0$ for all $i \geq 2$.

Lemma 1. *The probability that the best arm under context c is eliminated from context c on round r is at most*

$$2\, \exp\Big(-\tfrac{T}{2|\mathcal{P}|H_1^c R}\Big).$$

Proof. Fix a context c and recall the above convention that $a^{c,1}$ denotes an optimal arm for c. Assume $a^{c,1}$ was not eliminated before round r. Let $t_{r,i}$ denote the number of i.i.d. samples used to estimate $Q^\alpha(c, a^{c,i})$ in round r, and define the (effective) harmonic mean

$$\mathrm{harmonic}(u, v) \ := \ \big(u^{-1} + v^{-1}\big)^{-1}.$$

Then, by Hoeffding's inequality, for any arm $a^{c,i} \in \mathcal{A}_c^{(r)}$,

$$\Pr\big[\hat{Q}^{\alpha,(r)}(c, a^{c,1}) < \hat{Q}^{\alpha,(r)}(c, a^{c,i})\big] \leq \exp\Big(-\tfrac{1}{2}\ \mathrm{harmonic}(t_{r,1}, t_{r,i})\, \Delta^2_{c,a^{c,i}}\Big).$$

Let N_r be the number of arms in $\mathcal{A}_c^{(r)}$ whose empirical estimate exceeds that of $a^{c,1}$. Then

$$\begin{aligned}
\mathbb{E}[N_r] &\leq \sum_{a^{c,i} \in \mathcal{A}_c^{(r)}} \Pr\big[\hat{Q}^{\alpha,(r)}(c, a^{c,1}) < \hat{Q}^{\alpha,(r)}(c, a^{c,i})\big] \\
&\leq \sum_{a^{c,i} \in \mathcal{A}_c^{(r)}} \exp\Big(-\tfrac{1}{2}\ \mathrm{harmonic}(t_{r,1}, t_{r,i})\, \Delta^2_{c,a^{c,i}}\Big) \\
&\leq \sum_{a^{c,i} \in \mathcal{A}_c^{(r)}} \exp\Big(-\Delta^2_{c,a^{c,i}} \cdot \tfrac{T}{2|\mathcal{P}|\, \bar{N}_i\, \log_2 |\mathcal{A}|}\Big) \\
&\leq |\mathcal{A}_c^{(r)}| \max_{a^{c,i} \in \mathcal{A}_c^{(r)}} \exp\Big(-\Delta^2_{c,a^{c,i}} \cdot \tfrac{T}{2|\mathcal{P}|\, \bar{N}_{\max}\, \log_2 |\mathcal{A}|}\Big) \\
&\leq |\mathcal{A}_c^{(r)}|\ \exp\Big(-\tfrac{T}{2|\mathcal{P}|\, H_1^c\, R}\Big),
\end{aligned}$$

where $\bar{N}_i := \frac{N(a^{c,1})+N(a^{c,i})}{2} \leq \bar{N}_{\max}$ and we used $R = \lceil \log_2 |\mathcal{A}| \rceil$.

For the best arm to be eliminated in round r, at least half of the active arms must (incorrectly) rank above it, i.e., $N_r \geq \frac{1}{2}\,|\mathcal{A}_c^{(r)}|$. By Markov's inequality,

$$\Pr\Big[N_r \geq \tfrac{1}{2}|\mathcal{A}_c^{(r)}|\Big] \ \leq \ \frac{2\,\mathbb{E}[N_r]}{|\mathcal{A}_c^{(r)}|} \ \leq \ 2\, \exp\Big(-\tfrac{T}{2|\mathcal{P}|H_1^c R}\Big),$$

which proves the lemma. □

Lemma 2. *The probability that the best arm under context c is eliminated from context c on round r is at most*

$$3\, \exp\Big(-\tfrac{T}{8|\mathcal{C}|H_2^c R}\Big).$$

Proof. This follows by adapting Lemma 4.3 of Karnin, Koren, and Somekh (2013) to our setting (using the same indexing convention as above). In particular, the key intermediate step is that

$$\begin{aligned}
\mathbb{E}[N_r] &\leq \sum_{a^{c,i} \in \mathcal{A}_c^{(r)}} \Pr\big[\hat{Q}^{\alpha,(r)}(c, a^{c,1}) < \hat{Q}^{\alpha,(r)}(c, a^{c,i})\big] \\
&\leq \sum_{a^{c,i} \in \mathcal{A}_c^{(r)}} \exp\Big(-\Delta^2_{c,a^{c,i}} \cdot \tfrac{2^r T}{8|\mathcal{C}|\, |\mathcal{A}|\, \bar{N}_i\, \log_2 |\mathcal{A}|}\Big),
\end{aligned}$$

and proceeding exactly as in Karnin, Koren, and Somekh (2013) yields the stated bound (with complexity term H_2^c). □

Proof of Theorem 1. The best arm needs to survive for all R rounds and under all contexts $\mathcal{C}$. Therefore, from the Lemma 1:

$$\sum_{r=1}^{R}\sum_{c} 2\exp\Big(-\tfrac{T}{2|\mathcal{P}|H_1^c R}\Big) \;\le\; 3|\mathcal{C}|R\exp\Big(-\tfrac{T}{2|\mathcal{P}|H_1 R}\Big)$$

From the Lemma 2:

$$\sum_{r=1}^{R}\sum_{c} 3\exp\Big(-\tfrac{T}{8|\mathcal{C}|H_2^c R}\Big) \;\le\; 3|\mathcal{C}|R\exp\Big(-\tfrac{T}{8|\mathcal{C}|H_2 R}\Big)$$

Combining both:

$$3|\mathcal{C}|R\exp\Big(-\tfrac{T}{\min(2|\mathcal{P}|H_1, 8|\mathcal{C}|H_2)R}\Big)$$

which gives the theorem. □

Proposition 2 (Inference-Agnostic Optimality)**.** *The Inference-Agnostic prompt-optimization policy remains optimal under linear transformation of $R_x^{\mathrm{IA}}(c,a)$, that is, $\sigma R_x^{\mathrm{IA}}(c,a), \sigma > 0$ and an optimal policy can be recovered trivially from Q-function under affine transformation:*

$$Q^{AF}(c,a) := \mathbb{E}_{x\sim\mathcal{X}}\big[\sigma R_x^{\mathrm{IA}}(c,a)+\mu\big] = \sigma Q^{\mathrm{IA}}(c,a)+\mu.$$

Proof. By linearity of expectation,

$$\begin{aligned} Q'(c,a) &= \mathbb{E}_x\big[\sigma R_x^{\mathrm{IA}}(c,a)+\mu\big] \\ &= \sigma\,\mathbb{E}_x\big[R_x^{\mathrm{IA}}(c,a)\big]+\mu \\ &= \sigma Q^{\mathrm{IA}}(c,a)+\mu. \end{aligned}$$

Since $\sigma > 0$, for any two arms a, b we have

$$\begin{aligned} Q'(c,a) \ge Q'(c,b) &\iff \sigma Q^{\mathrm{IA}}(c,a)+\mu \ge \sigma Q^{\mathrm{IA}}(c,b)+\mu \\ &\iff Q^{\mathrm{IA}}(c,a) \ge Q^{\mathrm{IA}}(c,b). \end{aligned}$$

Therefore the ordering of arms is preserved and the $\arg\max$ set is identical. □

Appendix B

Synthetic-Bernoulli Environment. We consider a setting with $|\mathcal{P}| = 32$ prompts, each evaluated over a hidden mixture of query difficulty tiers—{easy, medium, hard}—spanning $|\mathcal{X}| = 520$ queries, with proportions $6:4:3$. For each prompt p and query x, the single-shot success probability is denoted $q_p(x) \in [0,1]$.

A pull of $N \le N_{\max}$ for prompt p on query x generates i.i.d. Bernoulli outcomes $\{z_i\}_{i=1}^{N}$ where $\Pr(z_i = 1) = q_p(x)$, and each completion incurs a per-completion cost k_p. The result is an array $\big[\, z_i,\ k_p \,\big]_{i=1}^{N}$.

Majority Voting (MV) sets $M = 1$ if $\sum_i z_i > N/2$, $M = 0$ if $\sum_i z_i < N/2$, and assigns $M = 0.5$ (by fair coin) in the case of a tie (N even, $\sum_i z_i = N/2$).

The cost-adjusted utility for context $c \in \{\text{low}, \text{mid}, \text{high}\}$ is computed as

$$u_c = w_1 M - w_2(c)\sum_{i=1}^{N} k_p,$$

where $w_1 = 1$ and $w_2(c) \in \{0,\ 0.2,\ 1.0\}$ depending on the cost tier.

To instantiate the environment, we generate two prompt archetypes: *deceiving prompts*, which achieve high average accuracy but exhibit low $q_p(x)$ on hard queries, and *all-rounders*, which maintain moderate accuracy more uniformly across tiers. Per-prompt costs k_p are sampled from a normal distribution with mean 0.02 and variance 0.005.

Synthetic-Categorical Environment. We model $|\mathcal{P}| = 32$ prompts, each paired with $|\mathcal{X}| = 512$ queries and $K = 2$ positive objectives. For every (p,x), there are M categorical outcomes, each represented by a vector $o_j \in \mathbb{R}^K$. A pull of $N \le N_{\max}$ (with $N_{\max} = 32$) for prompt p on query x generates N i.i.d. outcome vectors, resulting in rows $\big[\, o_{i,1},\ o_{i,2},\ k_p \,\big]$, where k_p denotes the per-completion cost for prompt p.

Given a context c with weights $w = (w_1, w_2, w_{\text{cost}})$, where $w_1 + w_2 = 1$ and $w_{\text{cost}} \ge 0$, the Best-of-N (BoN) utility is defined as

$$u_c = \max_{1\le i\le N}\,(w_1 o_{i,1} + w_2 o_{i,2}) - w_{\text{cost}}\, N\, k_p.$$

To construct the environment, outcome vectors are sampled from $\{-4,\dots,4\}^2$. We instantiate two prompt archetypes: ***HMLV*** (high mean, low variance; excels at $N{=}1$) and ***LMHV*** (lower mean, high variance; benefits from larger N), each specializing in one objective. For each (p,x), we add small per-query noise to the categorical outcome probabilities, introduce a mild train-to-test shift by perturbing these probabilities, sample per-prompt costs $k_p \in [0.02,\ 0.1]$, and draw context weights from a grid satisfying $w_1 + w_2 = 1$ with $w_{\text{cost}} \in \{0.1,\ 0.5,\ 1.0\}$.

MATH Environment. We select 316 integer-answer problems from the MATH dataset[1]. A set of 25 prompt templates is authored using *ChatGPT-o3*. For each (prompt, problem) pair, we sample 128 responses from `Llama-3.3-70B-Instruct` at temperature $T = 0.7$, parsing each completion to its final integer answer.

The dataset is then processed as follows:

1. For each problem, retain the global top-4 answers and group all other answers into a single OTHER bucket (five categories in total).
2. Compute per-prompt costs as the normalized average token length of its responses.

This yields a categorical environment (analogous to the Synthetic-Categorical setting) with $|\mathcal{P}| = 25$, $N_{\max} = 32$, a uniform context prior $c \in \{\text{low}, \text{mid}, \text{high}\}$, and cost coefficients $\{0,\ 0.2,\ 1.0\}$. Utility is evaluated via MV.

[1] https://huggingface.co/datasets/HuggingFaceH4/MATH-500

CommonsenseQA Environment. We randomly sample 1,500 multiple-choice questions from the CommonsenseQA corpus[2], and author 48 prompt templates using *ChatGPT-o3*. For each (prompt, question) pair, we query `Llama-3.3-70B-Instruct` at temperature $T = 1.1$, collecting 128 JSON-constrained answers (one of "Option A"–"Option E"). Each prompt is assigned a constant cost $k_p = 0.01$ as we use JSON mode, and all completion has equal length.

The resulting data is used to construct a categorical environment (in analogy to the Synthetic-Categorical setting) with $|\mathcal{P}| = 48$, $N_{\max} = 32$, a uniform context prior, and cost coefficients $\{0,\ 0.2,\ 1.0\}$.

Helpful–Harmless Environment. We filter the HH-RLHF conversations[3] to the 1,355 examples containing a single user query and a single assistant response. Using *ChatGPT-o3*, we craft 20 prompt templates. For each (prompt, query) pair, we sample 128 continuations from `Llama-3.3-70B-Instruct` at temperature $T = 0.7$. Each continuation is scored by separate public reward models (Yang et al. 2024) for *helpfulness*[4] and *harmlessness*[5], with scores normalized to $[-1, 1]$.

The two reward scores are then binned on a 0.5-spaced grid, producing a categorical distribution per (prompt, query); per-prompt costs are computed as the average token length. This data defines a categorical environment with $|\mathcal{P}| = 20$, $N_{\max} = 32$, and a uniform context prior over weight triples $(w_{\text{help}}, w_{\text{harm}}, w_{\text{cost}})$ with $w_{\text{help}} + w_{\text{harm}} = 1$ and $w_{\text{cost}} \in \{0.1,\ 0.5,\ 1.0\}$.

Summarization Environment. We randomly sample 1,201 Reddit posts from the Summarize-from-Feedback corpus[6] and design 20 summarization prompt templates using *ChatGPT-o3*. For each (prompt, post) pair, we query `Llama-3.3-70B-Instruct` at temperature $T = 0.7$ and collect 128 candidate summaries.

Each summary is scored by two publicly available reward models: *Preference*[7] and *Faithful*[8], with raw scores normalized to $[-1, 1]$. We then bin each dimension in steps of 0.5, producing a categorical distribution over the two reward dimensions, and compute per-prompt costs from average token length.

This data defines a categorical environment with $|\mathcal{P}| = 20$, $N_{\max} = 32$, and a uniform context prior over weight triples $(w_{\text{pref}}, w_{\text{faith}}, w_{\text{cost}})$ where $w_{\text{pref}} + w_{\text{faith}} = 1$ and $w_{\text{cost}} \in \{0.1,\ 0.5,\ 1.0\}$.

Note: All prompts are available under the prompts folder of the code base.

[2]https://huggingface.co/datasets/tau/commonsense_qa

[3]https://huggingface.co/datasets/Anthropic/hh-rlhf

[4]Ray2333/gpt2-large-helpful-reward_model

[5]Ray2333/gpt2-large-harmless-reward_model

[6]https://huggingface.co/datasets/openai/summarize_from_feedback

[7]OpenAssistant/reward-model-deberta-v3-large-v2

[8]CogComp/bart-faithful-summary-detector

Appendix C

Top-K screening. For the screening variant, we fixed $K = 4$ candidates after screening and swept the burn-in fraction $\rho \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$, which allocates a ρ-fraction of the budget to obtain initial estimates before trimming. The parameter-sweep protocol matched the baselines. We selected $\rho = 0.20$ for reporting, as it achieved the best overall performance while remaining robust across datasets and inference regimes (Table 1).

UCB. We tuned the exploration constant over $c \in \{0.1, 0.5, 1.0, 2.0, 4.0, 8.0\}$ under the same budgets, using 20% of the data per environment with identical seeds across settings, and 10,000 test contexts. The agent ranks arms by the standard UCB index

$$\mathrm{UCB}_i(t) \;=\; \hat{\mu}_i(t) \;+\; c\sqrt{\frac{\ln t}{n_i(t)}},$$

where $\hat{\mu}_i(t)$ is the empirical mean utility of arm i, $n_i(t)$ its pull count, and t the total pulls. We selected $c = 0.1$ for reporting, as it achieved the best overall performance while remaining robust across datasets and inference regimes (Table 2).

ϵ-greedy. We swept the greedy probability $1 - \epsilon \in \{0.50, 0.75, 0.80, 0.85, 0.90, 0.95\}$ separately for each dataset and inference regime (MV, BoN). For every ϵ, agents were trained under budgets $T \in \{3\text{K}, 5\text{K}, 10\text{K}, 20\text{K}, 30\text{K}, 40\text{K}\}$, using 20% of the data per environment with deterministic reseeding; evaluation used 10,000 test contexts per environment. We selected $\epsilon = 0.15$ for reporting, as it achieved the best overall performance while remaining robust across datasets and inference regimes (Table 3).

Appendix D

Statistical testing. For each dataset and budget T, we perform all pairwise algorithm comparisons using per-environment utilities as *paired* samples (identical train/test splits via deterministic reseeding). Our default test is the two-sided Wilcoxon signed-rank test, which we apply to the aligned vectors after removing non-finite values and dropping exact ties (`zero_method=wilcox`, `mode=auto`); pairs with fewer than two effective samples are skipped. When requested, we also report the paired sign test (binomial test on the sign of differences) after removing ties. To control multiplicity within each $(\text{dataset}, T)$ grid, we use Holm–Bonferroni adjustment by default (with options for Benjamini–Hochberg FDR or no correction). We declare a *winner* if the adjusted $p < \alpha = 0.05$; the direction is determined by the sign of the median difference $\mathrm{median}(x - y)$. In case of unequal environment counts across algorithms, samples are truncated to the minimum length to preserve pairing. Figures visualize the outcome matrix with entries in $\{-1, 0, +1\}$ indicating row-algorithm loss, non-significance, or win against the column algorithm, respectively.

All the results are shown in Figs 1, 2, 3, 4, 5, 6. Across all six datasets, we observe that PSST and the Top-K screening

Param × T	HH	Summarization	SC	SB	MATH	CQA
ρ=0.05, $T = 3000$	0.40 ± 0.00	0.20 ± 0.00	2.77 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
ρ=0.05, $T = 5000$	0.40 ± 0.00	0.22 ± 0.00	2.83 ± 0.02	0.83 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.05, $T = 10000$	0.42 ± 0.00	0.21 ± 0.00	2.83 ± 0.02	0.85 ± 0.01	0.81 ± 0.01	0.76 ± 0.00
ρ=0.05, $T = 20000$	0.43 ± 0.00	0.22 ± 0.00	2.87 ± 0.02	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.05, $T = 30000$	0.44 ± 0.00	0.23 ± 0.00	2.88 ± 0.01	0.84 ± 0.00	0.82 ± 0.01	0.77 ± 0.00
ρ=0.05, $T = 40000$	0.43 ± 0.00	0.23 ± 0.00	2.87 ± 0.02	0.84 ± 0.00	0.81 ± 0.00	0.77 ± 0.00
ρ=0.10, $T = 3000$	0.41 ± 0.00	0.21 ± 0.00	2.79 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.10, $T = 5000$	0.41 ± 0.00	0.22 ± 0.00	2.84 ± 0.02	0.84 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.10, $T = 10000$	0.42 ± 0.00	0.21 ± 0.00	2.86 ± 0.02	0.84 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.10, $T = 20000$	0.43 ± 0.00	0.23 ± 0.00	2.88 ± 0.02	0.84 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.10, $T = 30000$	0.44 ± 0.00	0.23 ± 0.00	2.89 ± 0.02	0.84 ± 0.00	0.81 ± 0.01	0.77 ± 0.01
ρ=0.10, $T = 40000$	0.44 ± 0.00	0.23 ± 0.00	2.88 ± 0.02	0.84 ± 0.00	0.82 ± 0.00	0.77 ± 0.00
ρ=0.20, $T = 3000$	0.41 ± 0.00	0.20 ± 0.00	2.77 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.20, $T = 5000$	0.41 ± 0.00	0.22 ± 0.00	2.84 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.20, $T = 10000$	0.43 ± 0.00	0.22 ± 0.00	2.85 ± 0.02	0.83 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.20, $T = 20000$	0.43 ± 0.00	0.23 ± 0.00	2.87 ± 0.01	0.84 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.20, $T = 30000$	0.44 ± 0.00	0.23 ± 0.00	2.89 ± 0.02	0.84 ± 0.00	0.82 ± 0.01	0.76 ± 0.00
ρ=0.20, $T = 40000$	0.44 ± 0.00	0.22 ± 0.00	2.88 ± 0.02	0.84 ± 0.00	0.82 ± 0.01	0.77 ± 0.00
ρ=0.30, $T = 3000$	0.41 ± 0.00	0.21 ± 0.00	2.81 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.30, $T = 5000$	0.41 ± 0.00	0.22 ± 0.00	2.85 ± 0.01	0.83 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.30, $T = 10000$	0.42 ± 0.00	0.22 ± 0.00	2.85 ± 0.02	0.84 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.30, $T = 20000$	0.43 ± 0.00	0.22 ± 0.00	2.88 ± 0.01	0.83 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
ρ=0.30, $T = 30000$	0.44 ± 0.00	0.22 ± 0.00	2.88 ± 0.01	0.84 ± 0.00	0.82 ± 0.01	0.76 ± 0.00
ρ=0.30, $T = 40000$	0.44 ± 0.00	0.23 ± 0.00	2.89 ± 0.01	0.84 ± 0.00	0.82 ± 0.01	0.77 ± 0.00
ρ=0.40, $T = 3000$	0.41 ± 0.00	0.20 ± 0.00	2.80 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.40, $T = 5000$	0.42 ± 0.00	0.20 ± 0.00	2.80 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
ρ=0.40, $T = 10000$	0.43 ± 0.00	0.22 ± 0.00	2.85 ± 0.01	0.83 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.40, $T = 20000$	0.43 ± 0.00	0.22 ± 0.00	2.87 ± 0.02	0.83 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.40, $T = 30000$	0.44 ± 0.00	0.22 ± 0.00	2.88 ± 0.01	0.83 ± 0.00	0.81 ± 0.01	0.77 ± 0.00
ρ=0.40, $T = 40000$	0.44 ± 0.00	0.23 ± 0.00	2.89 ± 0.01	0.84 ± 0.00	0.82 ± 0.01	0.77 ± 0.00

Table 1: PSST+K4: mean $\pm$ SEM across datasets (rows are param, ρ and T).

heuristic consistently outperform competing methods across most budget settings, with statistical significance.

References

Karnin, Z. S.; Koren, T.; and Somekh, O. 2013. Almost optimal exploration in multi-armed bandits. In *Proceedings of the the 30th International Conference on Machine Learning*.

Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment.

Param × T	HH	Summarization	SC	SB	MATH	CQA
c=0.1, $T = 3000$	0.38 ± 0.00	0.19 ± 0.01	2.83 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=0.1, $T = 5000$	0.39 ± 0.00	0.21 ± 0.00	2.88 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=0.1, $T = 10000$	0.41 ± 0.00	0.23 ± 0.00	2.94 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=0.1, $T = 20000$	0.43 ± 0.00	0.25 ± 0.00	2.98 ± 0.01	0.86 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=0.1, $T = 30000$	0.43 ± 0.00	0.25 ± 0.00	2.99 ± 0.01	0.88 ± 0.00	0.81 ± 0.01	0.76 ± 0.01
c=0.1, $T = 40000$	0.44 ± 0.00	0.25 ± 0.00	3.00 ± 0.01	0.89 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
c=0.5, $T = 3000$	0.37 ± 0.00	0.19 ± 0.01	2.85 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=0.5, $T = 5000$	0.38 ± 0.00	0.20 ± 0.00	2.90 ± 0.01	0.82 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
c=0.5, $T = 10000$	0.41 ± 0.00	0.23 ± 0.00	2.94 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=0.5, $T = 20000$	0.43 ± 0.00	0.24 ± 0.00	2.98 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=0.5, $T = 30000$	0.43 ± 0.00	0.24 ± 0.00	3.00 ± 0.01	0.86 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=0.5, $T = 40000$	0.44 ± 0.00	0.25 ± 0.00	3.00 ± 0.01	0.88 ± 0.00	0.81 ± 0.01	0.75 ± 0.00
c=1.0, $T = 3000$	0.37 ± 0.00	0.19 ± 0.01	2.88 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=1.0, $T = 5000$	0.37 ± 0.00	0.19 ± 0.01	2.91 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
c=1.0, $T = 10000$	0.41 ± 0.00	0.22 ± 0.00	2.94 ± 0.01	0.83 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
c=1.0, $T = 20000$	0.42 ± 0.00	0.24 ± 0.00	2.98 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=1.0, $T = 30000$	0.43 ± 0.00	0.24 ± 0.00	3.00 ± 0.01	0.86 ± 0.01	0.81 ± 0.01	0.76 ± 0.00
c=1.0, $T = 40000$	0.43 ± 0.00	0.25 ± 0.00	3.00 ± 0.01	0.87 ± 0.00	0.81 ± 0.01	0.76 ± 0.01
c=2.0, $T = 3000$	0.37 ± 0.00	0.18 ± 0.01	2.86 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=2.0, $T = 5000$	0.38 ± 0.00	0.19 ± 0.01	2.93 ± 0.01	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
c=2.0, $T = 10000$	0.40 ± 0.00	0.23 ± 0.00	2.94 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=2.0, $T = 20000$	0.42 ± 0.00	0.24 ± 0.00	2.98 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=2.0, $T = 30000$	0.42 ± 0.00	0.24 ± 0.00	2.99 ± 0.01	0.86 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
c=2.0, $T = 40000$	0.43 ± 0.00	0.25 ± 0.00	2.99 ± 0.01	0.88 ± 0.00	0.81 ± 0.01	0.75 ± 0.00
c=4.0, $T = 3000$	0.37 ± 0.00	0.18 ± 0.01	2.85 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=4.0, $T = 5000$	0.37 ± 0.00	0.18 ± 0.00	2.91 ± 0.01	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
c=4.0, $T = 10000$	0.41 ± 0.00	0.22 ± 0.00	2.94 ± 0.01	0.84 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
c=4.0, $T = 20000$	0.42 ± 0.00	0.24 ± 0.00	2.97 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
c=4.0, $T = 30000$	0.42 ± 0.00	0.24 ± 0.00	2.99 ± 0.01	0.86 ± 0.00	0.81 ± 0.01	0.75 ± 0.00
c=4.0, $T = 40000$	0.43 ± 0.00	0.25 ± 0.00	3.00 ± 0.01	0.87 ± 0.00	0.81 ± 0.01	0.76 ± 0.00
c=8.0, $T = 3000$	0.37 ± 0.00	0.18 ± 0.01	2.86 ± 0.02	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
c=8.0, $T = 5000$	0.38 ± 0.00	0.19 ± 0.01	2.90 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
c=8.0, $T = 10000$	0.40 ± 0.00	0.22 ± 0.00	2.92 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=8.0, $T = 20000$	0.42 ± 0.00	0.24 ± 0.00	2.97 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
c=8.0, $T = 30000$	0.43 ± 0.00	0.24 ± 0.00	2.98 ± 0.01	0.86 ± 0.00	0.81 ± 0.01	0.75 ± 0.00
c=8.0, $T = 40000$	0.43 ± 0.00	0.25 ± 0.00	2.99 ± 0.01	0.87 ± 0.00	0.81 ± 0.01	0.76 ± 0.01

Table 2: UCB: mean $\pm$ SEM across datasets (rows are param, T).

Param × T	HH	Summarization	SC	SB	MATH	CQA
e=0.50, $T = 3000$	0.37 ± 0.00	0.17 ± 0.01	2.78 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
e=0.50, $T = 5000$	0.39 ± 0.00	0.20 ± 0.01	2.82 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
e=0.50, $T = 10000$	0.41 ± 0.00	0.21 ± 0.00	2.90 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.50, $T = 20000$	0.42 ± 0.00	0.23 ± 0.00	2.94 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.74 ± 0.00
e=0.50, $T = 30000$	0.43 ± 0.00	0.24 ± 0.00	2.97 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
e=0.50, $T = 40000$	0.43 ± 0.00	0.25 ± 0.00	2.98 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.01
e=0.75, $T = 3000$	0.38 ± 0.00	0.16 ± 0.01	2.75 ± 0.03	0.83 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.75, $T = 5000$	0.39 ± 0.00	0.17 ± 0.01	2.86 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
e=0.75, $T = 10000$	0.40 ± 0.00	0.20 ± 0.01	2.91 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
e=0.75, $T = 20000$	0.42 ± 0.00	0.23 ± 0.00	2.95 ± 0.01	0.84 ± 0.00	0.79 ± 0.00	0.74 ± 0.00
e=0.75, $T = 30000$	0.43 ± 0.00	0.23 ± 0.00	2.97 ± 0.01	0.85 ± 0.00	0.81 ± 0.01	0.75 ± 0.00
e=0.75, $T = 40000$	0.43 ± 0.00	0.24 ± 0.00	2.96 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.74 ± 0.00
e=0.80, $T = 3000$	0.38 ± 0.00	0.18 ± 0.01	2.83 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.80, $T = 5000$	0.39 ± 0.00	0.19 ± 0.00	2.86 ± 0.02	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.80, $T = 10000$	0.40 ± 0.00	0.19 ± 0.01	2.91 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.80, $T = 20000$	0.42 ± 0.00	0.23 ± 0.00	2.94 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.80, $T = 30000$	0.41 ± 0.00	0.23 ± 0.00	2.96 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.80, $T = 40000$	0.43 ± 0.00	0.24 ± 0.00	2.98 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.74 ± 0.00
e=0.85, $T = 3000$	0.38 ± 0.00	0.16 ± 0.01	2.72 ± 0.04	0.83 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
e=0.85, $T = 5000$	0.38 ± 0.00	0.17 ± 0.01	2.87 ± 0.01	0.83 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.85, $T = 10000$	0.40 ± 0.00	0.20 ± 0.00	2.90 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.85, $T = 20000$	0.41 ± 0.00	0.22 ± 0.00	2.95 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.85, $T = 30000$	0.42 ± 0.00	0.23 ± 0.01	2.95 ± 0.01	0.85 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.85, $T = 40000$	0.42 ± 0.00	0.24 ± 0.00	2.97 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.01
e=0.90, $T = 3000$	0.37 ± 0.00	0.17 ± 0.01	2.81 ± 0.03	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
e=0.90, $T = 5000$	0.38 ± 0.00	0.17 ± 0.01	2.87 ± 0.02	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
e=0.90, $T = 10000$	0.40 ± 0.00	0.19 ± 0.01	2.90 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.90, $T = 20000$	0.41 ± 0.00	0.22 ± 0.00	2.94 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.90, $T = 30000$	0.42 ± 0.00	0.23 ± 0.00	2.95 ± 0.01	0.85 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.90, $T = 40000$	0.42 ± 0.00	0.24 ± 0.00	2.97 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
e=0.95, $T = 3000$	0.37 ± 0.00	0.17 ± 0.01	2.76 ± 0.03	0.82 ± 0.00	0.78 ± 0.01	0.75 ± 0.00
e=0.95, $T = 5000$	0.38 ± 0.00	0.17 ± 0.01	2.86 ± 0.01	0.83 ± 0.00	0.79 ± 0.01	0.76 ± 0.00
e=0.95, $T = 10000$	0.39 ± 0.00	0.19 ± 0.01	2.92 ± 0.01	0.84 ± 0.00	0.80 ± 0.01	0.76 ± 0.00
e=0.95, $T = 20000$	0.41 ± 0.00	0.21 ± 0.00	2.95 ± 0.01	0.84 ± 0.00	0.79 ± 0.01	0.75 ± 0.00
e=0.95, $T = 30000$	0.42 ± 0.00	0.22 ± 0.00	2.95 ± 0.01	0.85 ± 0.00	0.80 ± 0.01	0.75 ± 0.00
e=0.95, $T = 40000$	0.41 ± 0.00	0.23 ± 0.00	2.95 ± 0.01	0.84 ± 0.00	0.79 ± 0.00	0.75 ± 0.00

Table 3: ϵ-greedy: mean $\pm$ SEM across datasets (rows are param, T).

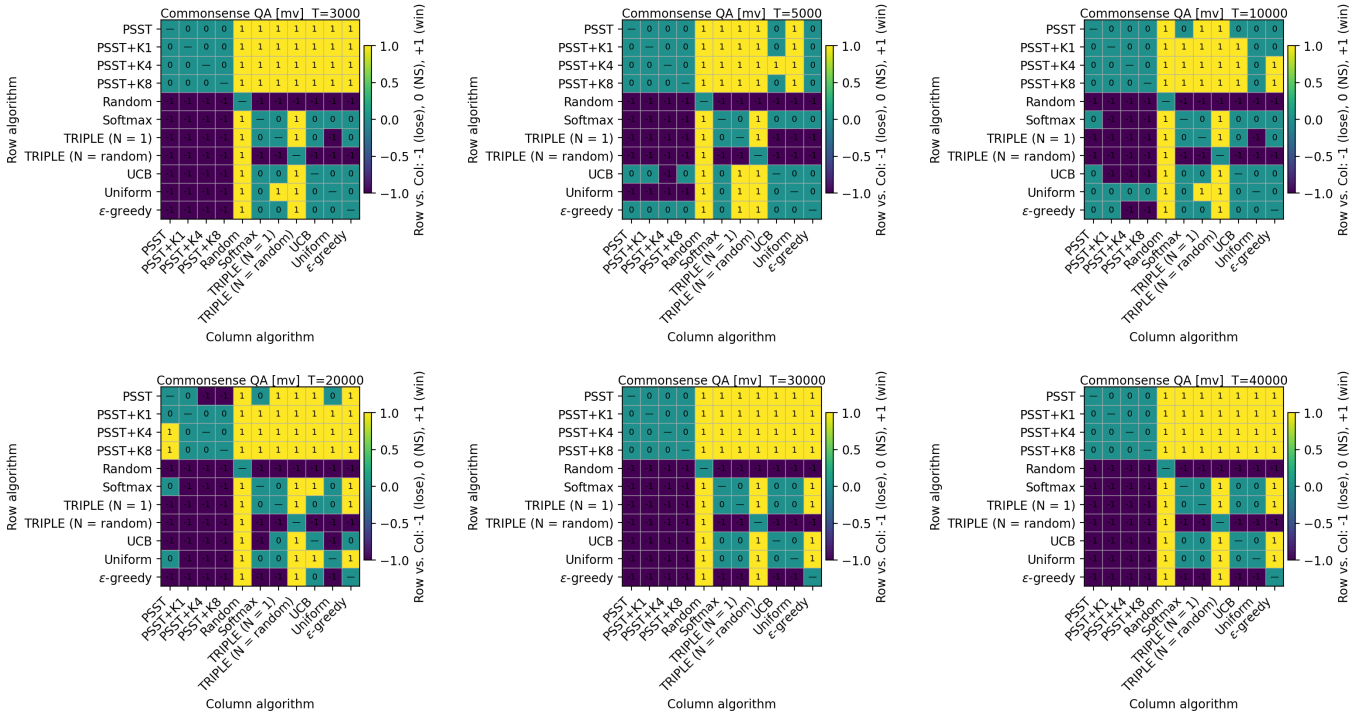

Figure 1: Pairwise wins for Commonsense QA (MV) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).

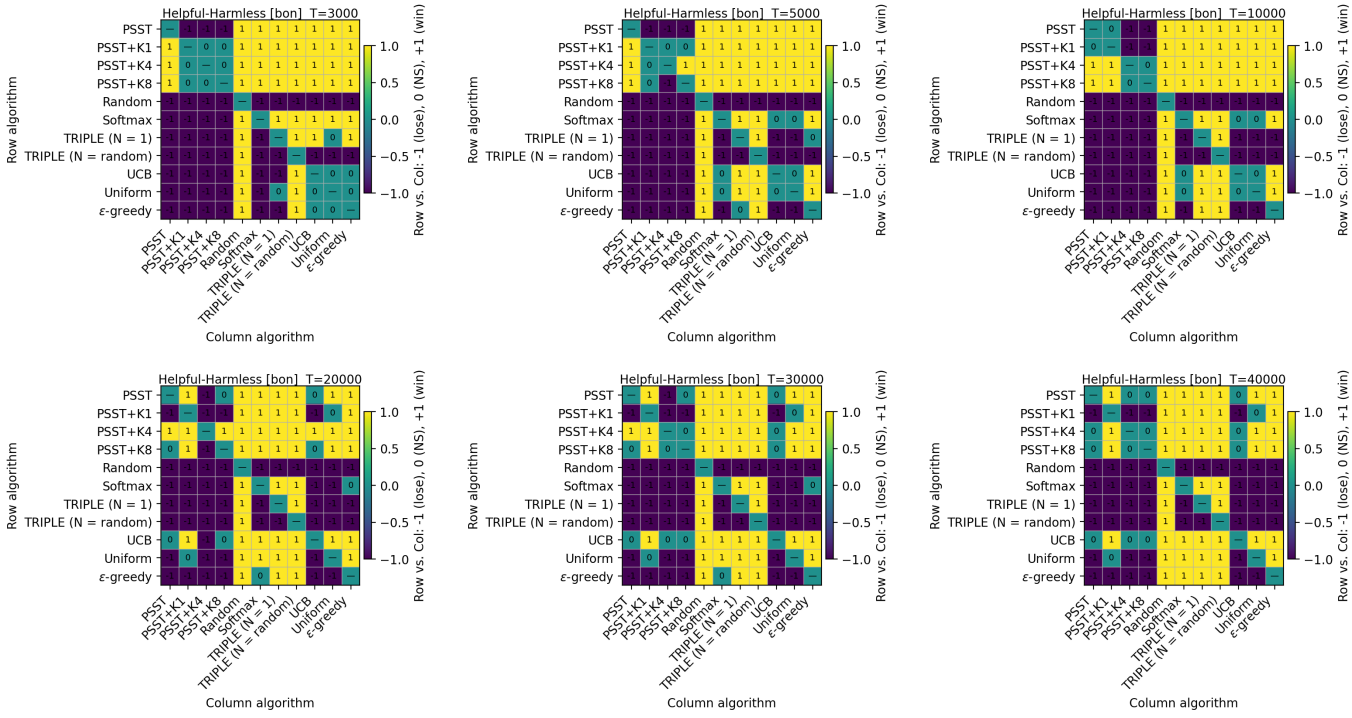

Figure 2: Pairwise wins for Helpful-Harmless (BoN) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).

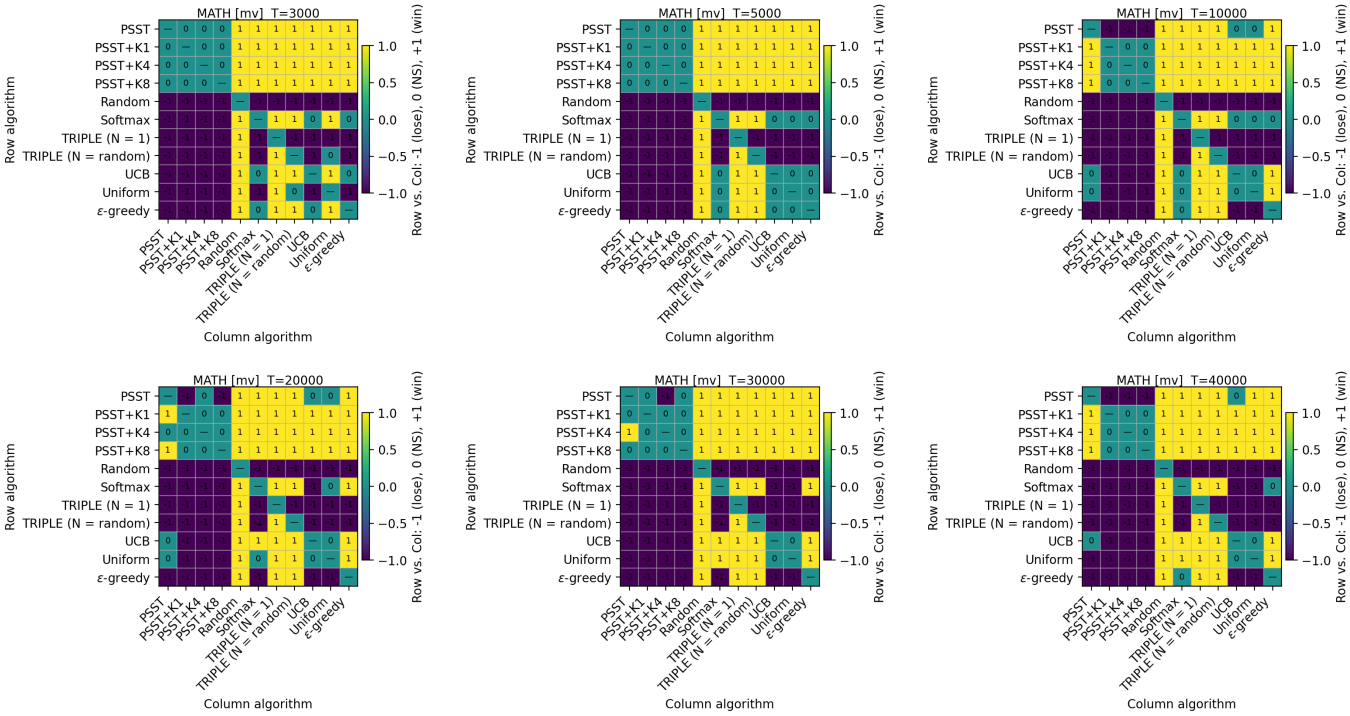

Figure 3: Pairwise wins for MATH (MV) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).

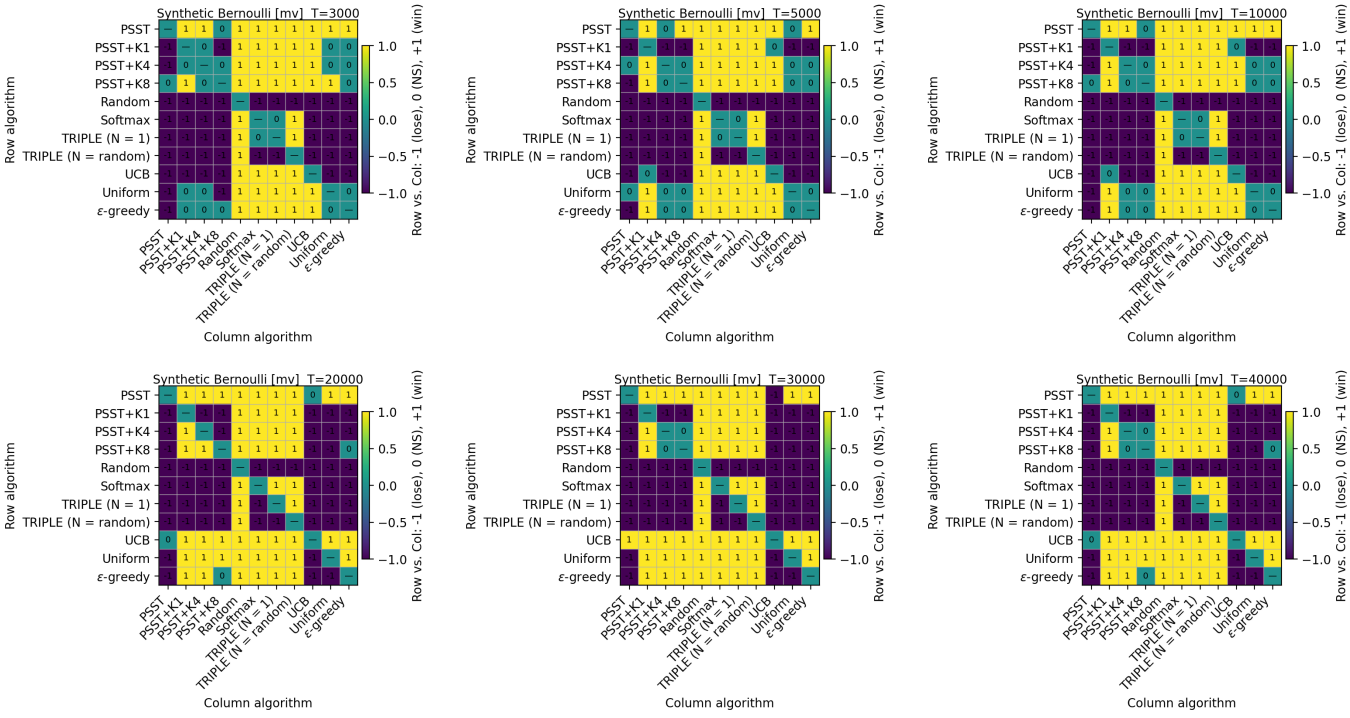

Figure 4: Pairwise wins for Synthetic Bernoulli (MV) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).

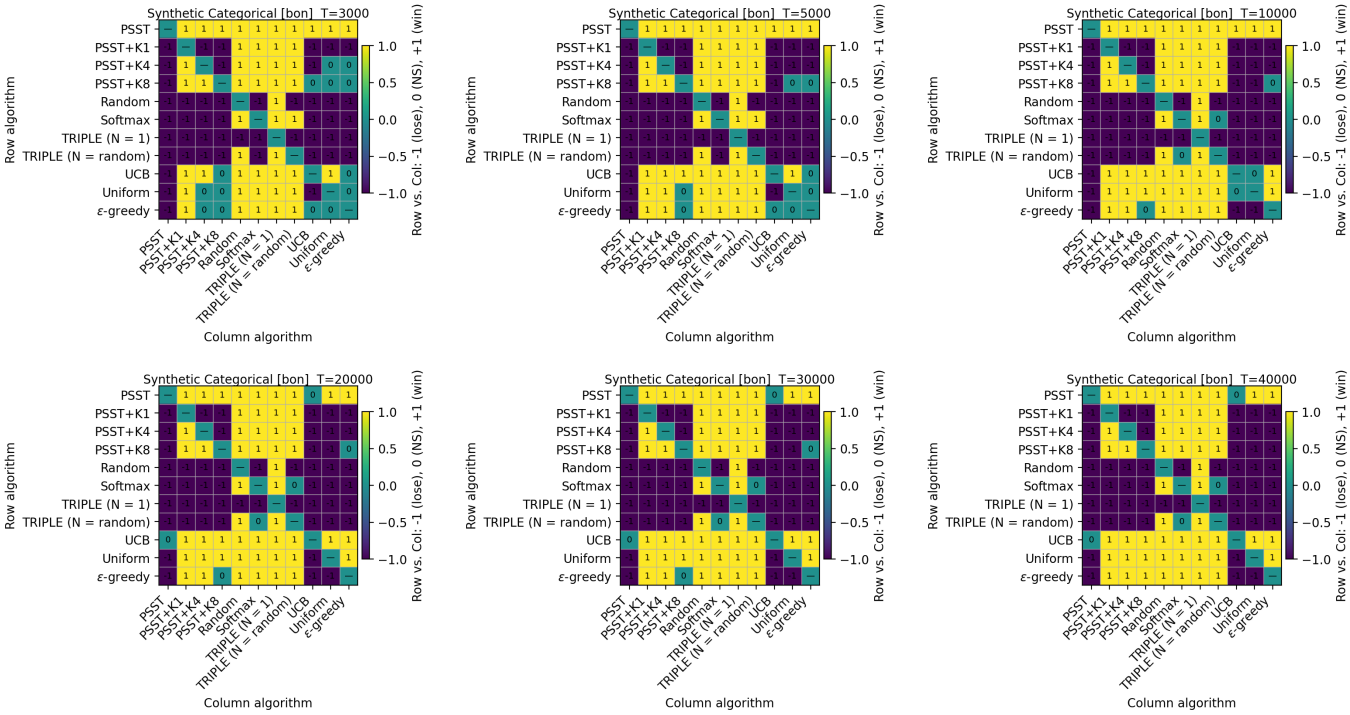

Figure 5: Pairwise wins for Synthetic Categorical (BoN) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).

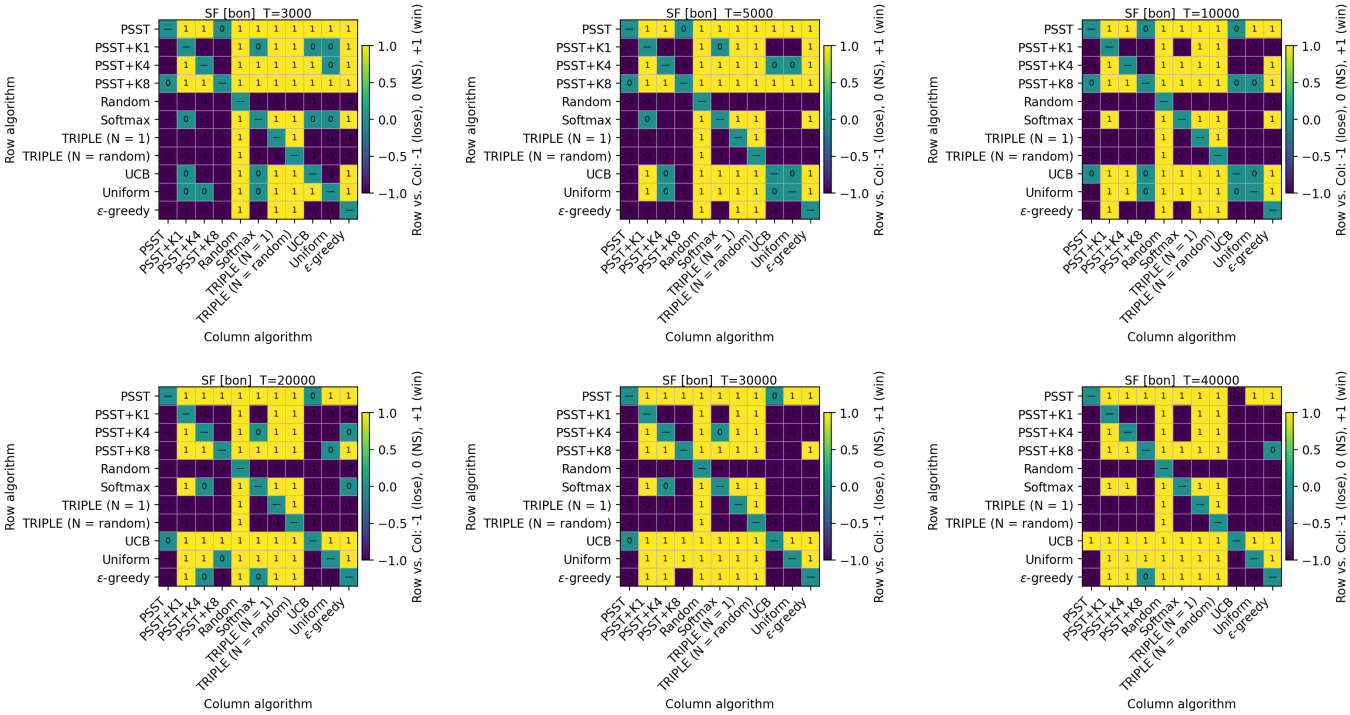

Figure 6: Pairwise wins for Summarization (BoN) across six budgets (T in order: 3000, 5000, 10000, 20000, 30000, 40000).