RELIABILITY-AWARE DETERMINANTAL POINT PROCESSES FOR ROBUST INFORMATIVE DATA SELECTION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional data sampling methods, including those based on Determinantal Point Processes (DPPs), offer deterministic approaches to maximize diversity, assuming that the selected data batches are always available without error. This presumption prohibits their use under probabilistic and disrupted access due to storage outage, imperfect communication, and hardware failures. This problem arises naturally in real-world LLM systems that aggregate data from distributed sources (e.g., RAG pipelines, teacher-student fine-tuning, and LLM-based cooperative perception and planning). We show that the original formulation of DPP-based inference collapses under such conditions. To address this gap, we introduce ProbDPP, a novel reliability-aware implementation of k-DPP that accounts for probabilistic data access by regularizing and decomposing the objective function into a geometric diversity term and unreliability cost, which remains well-posed, facilitating robust batch selection under uncertainty. To cover cases where failure probabilities are not known in advance, we frame this reliability-aware diversity maximization as a combinatorial semi-bandit problem and propose an algorithm that naturally adopts an efficient online learning policy. We evaluate the proposed method on different tasks, including question answering, RAG pipelines, summarization, and image reconstruction, across different benchmarks. Theoretical analysis provides regret bounds for the proposed approach under uncorrelated access links, ensuring performance guarantees.

1 INTRODUCTION

LLMs are increasingly being deployed across a wide range of applications. They rely on large amounts of curated data for training, fine-tuning, and teacher-student distillation. Recent works highlight that diversity-aware data selection can dramatically improve LLM performance and efficiency, often matching or surpassing full-data baselines while using only a fraction of the original dataset Yu et al. (2025); Wang et al. (2025); Jiang et al. (2025); Zhang et al. (2025a). Similarly, multimodal LLM inference is often constrained by prompt duration, especially when aggregating information from heterogeneous data sources Sun et al. (2025), urging elegant data selection and curation under strict context and computing budgets; a critical challenge in real-time multimodal applications such as LLM-based cooperative perception (CP) for autonomous driving Chiu et al. (2025); Wang et al. (2024a); Zhang et al. (2024); Sarlak et al. (2026). Compounding this bottleneck, upstream candidate data units are often unreliable Sarlak et al. (2023); Alinejad et al. (2024); Cho et al. (2024); Han et al. (2024), which can cause selected items to contribute negligible or degraded values to the downstream prompts or training pipelines Sarlak et al. (2025); Ye et al. (2025); Huang et al. (2025). While diversity-driven objectives like log-determinant maximization in Determinantal Point Processes (DPPs) effectively curb redundancy by selecting embeddings that span large semantic volumes Zhang et al. (2025b); Chen et al. (2025); Lamont et al. (2024); Hough et al. (2006); Taskar (2013); Wang et al. (2024b), they fundamentally rely on the assumption of deterministic data presence; this assumption breaks down in stochastic, multi-source pipelines where candidate units are prone to random corruption, storage failures, denial of service (DoS) attacks to data centers, or packet drops due to intermittent communications Kulesza & Taskar (2010).

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.

As shown in Figure 1, our method of jointly modeling diversity and reliability, ProbDPP, strictly dominates both diversity-only and reliability-only baselines across a comprehensive suite of robust reconstruction benchmarks.

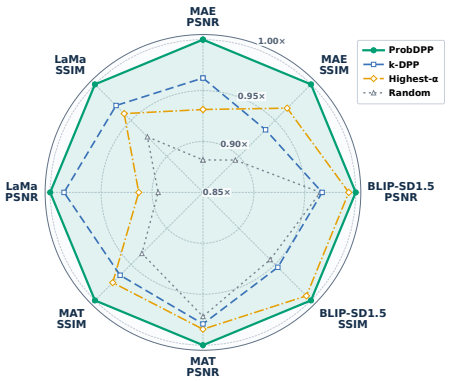


Figure 1: ProbDPP consistently dominates across 8 distinct multimodal and image reconstruction metrics (normalized). By jointly optimizing semantic coverage and transmission reliability, it outperforms competitors, including diversity-only (k-DPP) and reliability-only (Highest-α) baselines.

Direct optimization of the expected log-determinant, used in conventional DPP methods Kulesza & Taskar (2012), becomes ill-posed under Bernoulli dropouts, because the objective diverges to $-\infty$ whenever a data unit has a nonzero failure probability, exposing a fundamental gap between conventional diversity-based selection Neumann et al. (2021); Jha et al. (2024) and the stochastic reality of modern prompting pipelines Cuconasu et al. (2024). Figure 2 illustrates how a dropped unit makes the dropout-masked kernel singular, while Figure 3 illustrates the diversity-reliability trade-off addressed by ProbDPP. To overcome the ill-posedness of naive expected log-determinant diversity under Bernoulli dropouts, we introduce **ProbDPP**, a regularized subset selection framework that restores well-posedness and yields an exact, tractable decomposition by separating geometric embedding diversity from an additive per-unit reliability reward. While this objective enables principled data selection under known marginal availability, we extend the formulation to practical settings with unknown source reliabilities. The idea is casting online selection as a combinatorial semi-bandit with semi-bandit feedback Kveton et al. (2015); Wen et al. (2015) and designing an efficient KL-UCB-style algorithm Garivier & Cappé (2011) to dynamically balance diversity and reliability. Theoretically, we prove logarithmic regret upper bounds for our online policy under independent stationary failures and establish a complementary information-theoretic lower bound in a canonical singleton reduction Lattimore & Szepesvári (2020) to characterize the fundamental complexity and obtain performance guarantee for reliability-aware diversity maximization.

2 Related Work

Budgeted context selection in in-context learning (ICL), retrieval-augmented generation (RAG), and prompt compression increasingly relies on diversity-maximizing objectives—such as DPP and submodular volume maximization—to mitigate redundancy and maximize semantic coverage Wang et al. (2024c); Li et al. (2024); Jiang et al. (2023; 2024); Kulesza & Taskar (2011b;a; 2010); Taskar (2013), with recent work on formalizing black-box compression through rate-distortion limits Nagle et al. (2024). However, these frameworks assume that selected data units are deterministically available and usable. In multi-source and communication-constrained pipelines (e.g., edge-based RAG or cooperative driving), candidate units suffer from stochastic dropouts, storage failures, and noisy corruptions; under independent Bernoulli availability, the naive expected log-determinant objective of DPP collapses to $-\infty$, revealing an unaddressed gap in classical diversity-based selection. While robust and deletion-robust submodular maximization address worst-case removals or bounded adversarial perturbations Anari et al. (2019); Mirzasoleiman et al. (2017); Krause et al. (2008), they do not resolve the probabilistic singularity caused by stochastic Bernoulli masking. Similarly, diversity-only participant selection methods in federated learning (FL) apply DPP to profile non-IID client distributions Zhang et al. (2023); Bastola et al. (2024), but assume participation reliability is fixed and known a priori. Our method addresses this gap by assuming probabilistic and a priori known access. When source reliabilities are unknown and must be learned online, the combinatorial semi-bandit approach provides principled regret minimization Sankararaman (2016); Kveton et al. (2015); Garivier & Cappé (2011). Our problem departs from standard linear semi-bandit settings by coupling an unknown linear reliability reward with a known, nonlinear log-determinant diversity metric, bridging combinatorial bandit learning with robust diversity selection under stochastic data access.

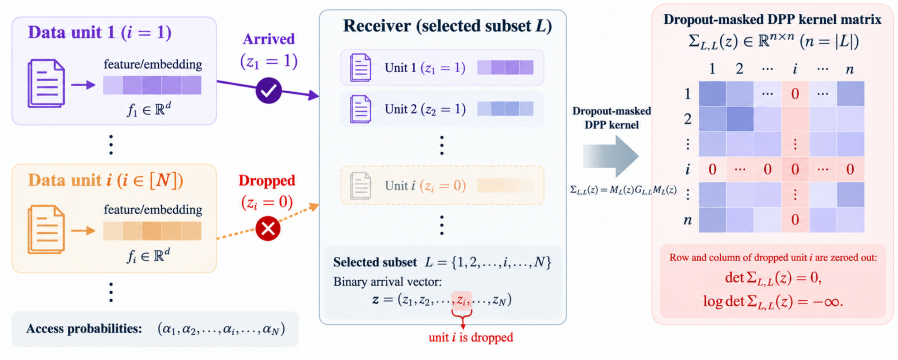


Figure 2: Ill-posedness of the expected dropout-masked log-determinant. For the fixed selected subset $L = \{a, b\}$, unit a is received and unit b is dropped. The corresponding zero row and column make $\Sigma_{L,L}(z)$ singular, yielding $\log\det\Sigma_{L,L}(z) = -\infty$. A positive probability of such a dropout makes the expected log-determinant equal to $-\infty$.

3 METHODOLOGY

We consider N candidate sources (clients) indexed by $[N] = \{1, \ldots, N\}$. At each round $t = 1, 2, \ldots, T$, the server selects a fixed-size subset $S_t \subseteq [N]$, $|S_t| = K$. Due to storage outage, imperfect communication, hardware failures, and stochastic unavailability, each selected source $i \in S_t$ succeeds with marginal probability $\alpha_i \in [0, 1]$. Let $z_{i,t} \sim \text{Bernoulli}(\alpha_i)$ denote the success indicator; the learner observes $\{z_{i,t} : i \in S_t\}$ only after attempting to fetch data (semi-bandit feedback). For each candidate source i, the server extracts and transmits a unit-norm feature embedding $f_i \in \mathbb{R}^d$ with $\|f_i\|_2 = 1$ (equivalently, the received feature at time t is $z_{i,t} f_i$). Similarity across received data embeddings is modeled by the positive semi-definite Gram matrix $G \in \mathbb{R}^{N \times N}$, $G_{ij} = \langle f_i, f_j \rangle$, $G \succeq 0$. Therefore, a standard diversity score for a selected subset $L \subseteq [N]$ is $g(L) \triangleq \log\det(G_{L,L})$.

Why the naïve dropout-aware log-det collapses. Under stochastic availability, the actually-received kernel for a chosen subset L is masked by the dropout indicators. Let $z_t = (z_{1,t}, \ldots, z_{N,t})$ be the binary mask vector and define the diagonal mask $M_L(z_t) = \text{diag}(z_{i,t} : i \in L)$. The unregularized received kernel is

$$\Sigma_{L,L}(z_t) = M_L(z_t)\, G_{L,L}\, M_L(z_t). \tag{1}$$

Figure 2 illustrates how dropouts make the unregularized kernel singular.

Lemma 1 (Ill-posedness under dropouts). *Fix any subset $L \subseteq [N]$ such that there exists $i \in L$ with $\alpha_i < 1$. Let $z_i \sim \text{Bernoulli}(\alpha_i)$ be independent and define $\Sigma_{L,L}(z) = M_L(z)\, G_{L,L}\, M_L(z)$ with $M_L(z) = \text{diag}(z_i : i \in L)$. Then*

$$\mathbb{E}_z\big[\log\det(\Sigma_{L,L}(z))\big] = -\infty. \tag{2}$$

A key insight is that a single selected item with non-zero failure probability is sufficient to make the dropout-masked kernel singular with positive probability, causing the expected log-determinant to collapse. *Proof.* See Appendix A.1.

3.1 RELIABILITY-AWARE PROBDPP OBJECTIVE

To solve the ill-posedness issue and obtain a finite and meaningful objective in dropout, we define $\epsilon > 0$ as a slack variable and adopt a regularized ProbDPP formulation whose masked kernel is:

$$W_L(z_t) = M_L(z_t) + \epsilon I, \quad \Sigma^{(\epsilon)}_{L,L}(z_t) = W_L(z_t)\, G_{L,L}\, W_L(z_t). \tag{3}$$

The resulting regularized expected diversity is $D_\epsilon(L) \triangleq \mathbb{E}_{z_t}\Big[\log\det\Big(\Sigma^{(\epsilon)}_{L,L}(z_t)\Big)\Big]$, which remains well posed for $\epsilon > 0$ and any choice of success probability α_i.

Lemma 2 (Exact decomposition of the regularized objective). *Let* $\epsilon > 0$ *and* $\Sigma^{(\epsilon)}_{L,L}(z) = W_L(z)\, G_{L,L}\, W_L(z)$ *with* $W_L(z) = M_L(z) + \epsilon I$. *Then*

$$D_\epsilon(L) \;=\; \log\det(G_{L,L}) + \sum_{i\in L} r_i(\alpha_i, \epsilon), \tag{4}$$

where the per-item reliability term is

$$r_i(\alpha, \epsilon) \;=\; 2\Big[\alpha \log(1+\epsilon) + (1-\alpha)\log\epsilon\Big]. \tag{5}$$

This exact decomposition is the central property of ProbDPP: it converts the stochastic log-determinant objective into a geometric diversity term plus an additive reliability reward, making reliability-aware subset selection tractable. *Proof.* See Appendix A.2.

Interpretation. For $G_{L,L} \succ 0$ and $\epsilon > 0$, the regularized kernel remains positive definite for every dropout realization. The resulting expected log-determinant decomposes into geometric diversity and an additive reliability contribution. Because $r_i(\alpha_i, \epsilon)$ increases with α_i, ProbDPP balances semantic diversity against the probability that selected information is available. More precisely, under fixed cardinality, ϵ induces an explicit reliability weight $\lambda_\epsilon = 2\log(1 + 1/\epsilon)$; our analysis shows that this weight decreases monotonically with ϵ and produces a closed-form switching threshold between the subsets favored by reliability and diversity (Appendix A.3).

Figure 3 illustrates this trade-off, where ProbDPP favors Sets 2 and 3, retaining complementary information while improving reliability relative. The precise selection depends on the Gram matrix of the embeddings and ϵ.

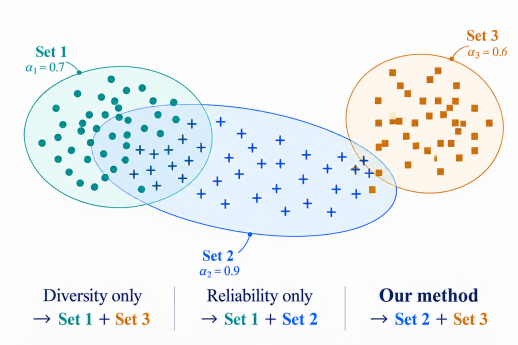


Figure 3: Conceptual illustration of the diversity–reliability trade-off. Sets 1 and 3 are selected by diversity methods for their lower semantic overlap; sets 1 and 2 are selected by reliability-only methods for their largest access probability; ProbDPP selects set 2 and set 3.

Online objective. When the reliabilities $\{\alpha_i\}_{i=1}^N$ are unknown in advance, the learner must select $\{S_t\}_{t=1}^T$ using only semi-bandit feedback $\{z_{i,t} : i \in S_t\}$. The goal is to maximize cumulative reliability-aware diversity via selecting S_t at time points $t = 1, 2, ..., T$ $\max_{|S_t|=K;\, S_t\subseteq[N]} \sum_{t=1}^T D_\epsilon(S_t)$, via gradual learning of success probability α_i from observed success indicators z_i.

Dependence structure. The decomposition in Lemma 2 depends only on the marginal success probabilities α_i and does not require independence across sources ($\alpha_i \perp \alpha_j$). Likewise, Eq. 4 depends only on $\Pr(z_i = 1) = \alpha_i$. Therefore, the offline ProbDPP objective remains valid under correlated or structured binary failures, as long as the marginal reliabilities are well defined. Furthermore, we report results for the temporally independent setting, $\mathbb{E}[z_{i,t_1} z_{i,t_2}] = \mathbb{E}[z_{i,t_1}]\,\mathbb{E}[z_{i,t_2}]$, which is the setting assumed in our regret analysis.

3.2 Online ProbDPP: A Combinatorial Semi-Bandit Formulation

The reliability-aware client selection problem naturally fits the combinatorial semi-bandit setting: at each round, the learner (at the source) selects multiple items under a cardinality constraint, and receives success indicator notifications via the feedback channel. We formalize the problem as a combinatorial semi-bandit with semi-bandit feedback, explicitly absorbing the unreliability parameters into the objective function. The bandit actions are modeled as a binary vector $A_t \in \{0,1\}^N$, $\|A_t\|_1 = K$ where $A_{t,i} = 1$ if and only if client i is selected and $A_{t,i} = 0$ otherwise. Specifically,

we define the induced linear coefficient $\theta_i \triangleq r_i(\alpha_i, \epsilon)$ where $r_i(\cdot, \epsilon)$ is a known reliability-to-reward mapping parameterized by ϵ, as defined in Equation 5. The unknown parameter vector is thus $\theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^N$. At each round t, the learner receives a linear reward contribution

$$\langle A_t, \theta \rangle = \sum_{i=1}^{N} A_{t,i}\, \theta_i = \sum_{i=1}^{N} A_{t,i}\, r_i(\alpha_i, \epsilon). \tag{6}$$

In addition to this unknown linear term, each action A_t is associated with a known deterministic combinatorial component $g(A_t) = \log\det\big(G_{S_t,S_t}\big)$, where G is a known Gram matrix and G_{S_t,S_t} denotes the principal submatrix indexed by the selected subset S_t. Combining both components, the per-round objective is to maximize $f(A_t) = g(A_t) + \langle A_t, \theta \rangle$. This formulation shows that the reliability-aware client selection problem induced by ProbDPP can be viewed as a combinatorial semi-bandit with a known nonlinear term and an unknown linear term. The learner estimates the Bernoulli parameters α_i from the feedback, constructs optimistic upper confidence bounds for these reliabilities, and maps them to optimistic coefficients θ_i for action selection.

After selecting an action $A_t \in \mathcal{A}$, the learner observes semi-bandit feedback: Let $n_i(t) = \sum_{s=1}^{t} \mathbb{I}\{A_{s,i} = 1\}$ denote the number of times coordinate i has been selected up to round t. The empirical mean estimator of the unknown reliability α_i is $\hat{\alpha}_i(t) = \frac{\sum_{s=1}^{t} \mathbb{I}\{A_{s,i}=1\}\, z_{i,s}}{\sum_{s=1}^{t} \mathbb{I}\{A_{s,i}=1\}}$ for $n_i(t) \geq 1$. This estimator coincides with the standard semi-bandit estimator for combinatorial bandits with deterministic action selection, when success estimates $\hat{\alpha}_i$ approach their target values α_i. No importance weighting or unbiased correction is required, since each observed outcome $z_{i,t}$ corresponds to an actual pull of coordinate i.

To balance exploration and exploitation, we construct optimistic upper confidence bounds for the unknown client reliabilities using the KL-UCB principle in Lattimore & Szepesvári (2020). Since the observed feedback variables $z_{i,t}$ are Bernoulli, an upper confidence bound for each reliability parameter α_i is defined as

$$U_i^{\alpha}(t) = \max\Big\{u \in [0,1] : n_i(t-1)\, \mathrm{kl}\big(\hat{P}_{\alpha_i(t)} \,\|\, P_u\big) \leq \log t + c \log\log t\Big\}, \tag{7}$$

where P_u represents a Bernoulli distribution with parameter u, and $\mathrm{kl}(P\|Q)$ denotes the Kullback-Leibler divergence between distributions P and Q. The constant $c > 0$ controls the confidence level. The optimistic reliability bounds are then mapped into optimistic linear contributions through the known reliability-to-reward transformation. Specifically, we define the optimistic coefficient $\tilde{\theta}_i(t) = r_i\big(U_i^{\alpha}(t), \epsilon\big)$, and collect these into the optimistic weight vector $\tilde{\theta}(t) = \big(\tilde{\theta}_1(t), \ldots, \tilde{\theta}_N(t)\big)$.

At each round t, the learner selects the action $A_t \in \arg\max_{a \in \mathcal{A}} \Big\{g(a) + \langle a, \tilde{\theta}(t)\rangle\Big\}$, where $g(a)$ and $\langle a, \tilde{\theta}(t)\rangle$ are defined respectively in 3.2 and equation 6 which corresponds to reliability-aware ProbDPP selection with coordinate-wise KL-UCB optimism applied to the induced linear coefficients. The regret guarantee assumes exact optimistic maximization. A greedy implementation requires a separate optimization-error analysis. The resulting online policy is tailored to ProbDPP's mixed structure: optimism is applied only to the unknown reliability coefficients, while the known nonlinear log-determinant term is optimized jointly with these optimistic rewards. The summary of this process is presented in **Algorithm 1** (see Appendix A.4).

4 Regret Formulation and Canonical Lower Bound

In this section, we establish regret guarantees for the proposed online learning framework. All proofs are deferred to the appendix unless stated otherwise. Actions depend only on past observations and independent algorithmic randomness. The feasible action family is finite, every action selects K sources, and $g(a)$ is known, fixed, and finite for every feasible action.

Let $A^\star \in \arg\max_{a \in \mathcal{A}} f(a)$ denote the best fixed size-K subset under the true reliabilities. For actions $\{A_t\}_{t=1}^{T}$ selected by ProbDPP, the cumulative pseudo-regret is $\mathrm{Reg}_T \triangleq \sum_{t=1}^{T}\big(f(A^\star) - f(A_t)\big)$. We assume stationary, independent Bernoulli semi-bandit feedback; the full algorithm, gap definitions, and proof are deferred to Appendices A.4 and A.5.

Table 1: Robustness under stochastic context unavailability on MeetingBank (max 30 chunks). Higher is better; best results are in bold and second-best results are underlined.

	LLMlingua2	LLMlingua-small	Random	ProbDPP
Token-F1	28.8	<u>29.5</u>	23.6	**31.3**
ROUGE-L	28.8	<u>29.4</u>	23.8	**31.2**
BERTScore	<u>32.5</u>	30.6	27.0	**36.9**
EM	<u>17.3</u>	16.9	15.0	**19.4**

Theorem 1 (Regret of ProbDPP). *Let* $\beta_\epsilon \triangleq 2\log\frac{1+\epsilon}{\epsilon}$ *and* $\alpha_i^+ \triangleq \min\Big\{1, \alpha_i + \frac{\Delta_i^{\min}}{K\beta_\epsilon}\Big\}$, *where* $\Delta_i^{\min}$ *and* $\Delta_i^{\max}$ *are the item-wise suboptimality gaps defined in Appendix A.5. Then*

$$\mathbb{E}[\mathrm{Reg}_T] \le \sum_{i=1}^{N} \Delta_i^{\max} \frac{\log T + c\log\log T}{D_{\mathrm{KL}}(P_{\alpha_i} \| P_{\alpha_i^+})} + O(1), \tag{8}$$

where $c > 0$ *is the KL-UCB exploration constant. Hence, ProbDPP achieves logarithmic expected pseudo-regret in the horizon* T. *More explicitly, if the item-dependent coefficients* $\frac{\Delta_i^{\max}}{D_{\mathrm{KL}}(P_{\alpha_i} \| P_{\alpha_i^+})}$ *are uniformly bounded across* i *independently of* N, *then equation 8 scales as* $O(N \log T)$ *in the number of sources* N *and horizon* T.

We complement the upper bound with a canonical logarithmic lower bound in the singleton-action reduction. Consider the special case $K = 1$ and a Gram matrix G such that $g(a)$ is constant over singleton actions. Then each action selects one item, and the objective reduces to a classical Bernoulli bandit with mean reward $r_i(\alpha_i, \epsilon)$. Since $r_i(\alpha, \epsilon)$ is affine and increasing in α, standard change-of-measure arguments imply that any uniformly good policy satisfies $\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{i \ne i^\star} \frac{\Delta_i}{\mathrm{kl}(\alpha_i \| \alpha_{i^\star})}$. For general K, Theorem 1 gives logarithmic regret with item-dependent gap constants; we do not claim a matching instance-dependent lower bound beyond this singleton reduction. The key reduction embeds a classical Bernoulli bandit as a special case of ProbDPP, showing that logarithmic dependence on T is unavoidable even in this canonical reliability-aware setting. The proof is given in Appendix A.6. A controlled stationary experiment matching the assumptions of our online analysis, including cumulative pseudo-regret and reliability-estimation error is provided in Appendix A.7.

5 Experiment Results

Robustness Under Stochastic Context Unavailability. We evaluate robustness to stochastic context unavailability on the MeetingBank QA benchmark Hu et al. (2023) using Llama-3. In our compress-then-prompt pipeline, we enforce a strict budget of 30 chunks (60–90 tokens each), which are subsequently subjected to independent Bernoulli dropouts ($\alpha \in [0, 0.8]$) at inference time. We compare ProbDPP against Random selection and state-of-the-art deterministic prompt compressors, LLMLingua2 Pan et al. (2024) and LLMLingua-small Jiang et al. (2023). Across 120 scenarios, ProbDPP achieves the highest scores on all four metrics (Table 1). Compared with the second-best method, ProbDPP improves Token-F1, ROUGE-L, BERTScore, and EM by 6.1%, 6.1%, 13.5%, and 12.1%, respectively. These results support incorporating availability alongside semantic diversity when selecting context under stochastic dropouts.

Robust Visual Transmission and Reconstruction. To demonstrate the versatility of our framework beyond textual data, we evaluate ProbDPP on visual token selection under stochastic access links using Masked Autoencoders (MAE). Under heterogeneous packet reliability, ProbDPP achieves the highest PSNR and SSIM among the evaluated methods (Table 2). Compared with Highest-α, it improves PSNR by 7.4% despite a 4.5% lower arrival rate. It also improves PSNR over diversity-only k-DPP by 3.9%. Figure 4 illustrates the corresponding qualitative differences. Additional qualitative and quantitative results for BLIP, LaMa, and MAT are provided in Appendix A.8.

Table 2: Visual reconstruction performance using Masked Autoencoders (MAE) under stochastic patch dropping. Arrows indicate the preferred direction; best results are bold and second-best results are underlined. For methods with fixed selection probabilities, the empirical arrival rates closely match their corresponding α values because they are averaged over a large number of patches.

Method	**Arrival Rate** ↑	**Full MSE** ↓	**PSNR** ↑	**SSIM** ↑
Random	0.70 ± 0.04	0.0170 ± 0.0002	19.41 ± 0.04	0.60 ± 0.00
Highest-α	$\mathbf{0.89 \pm 0.05}$	0.0111 ± 0.0006	20.50 ± 0.23	$\underline{0.64 \pm 0.01}$
k-DPP	0.63 ± 0.07	$\underline{0.0104 \pm 0.0002}$	$\underline{21.18 \pm 0.09}$	0.62 ± 0.00
CLS-Attn (ICLR'22)	0.65 ± 0.03	0.0152 ± 0.0002	19.34 ± 0.06	0.58 ± 0.00
FastV (ECCV'24)	0.66 ± 0.03	0.0152 ± 0.0002	19.43 ± 0.06	0.59 ± 0.00
Zero-TPrune (CVPR'24)	0.64 ± 0.05	0.0150 ± 0.0002	19.48 ± 0.06	0.59 ± 0.00
Self-Guided (NeurIPS'24)	0.67 ± 0.06	0.0210 ± 0.0002	17.75 ± 0.05	0.56 ± 0.00
ProbDPP (Ours)	$\underline{0.85 \pm 0.08}$	$\mathbf{0.0085 \pm 0.0003}$	$\mathbf{22.01 \pm 0.14}$	$\mathbf{0.67 \pm 0.00}$

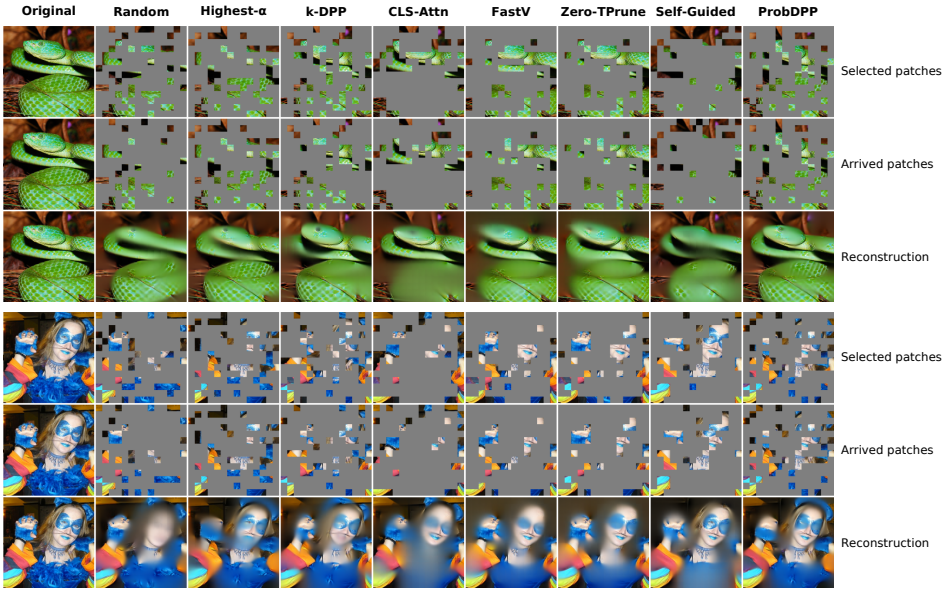


Figure 4: Qualitative comparison of MAE reconstruction under stochastic patch dropping. For each example, rows show selected patches, arrived patches, and the reconstructed image.

RAG context selection. Table 3 presents a RAG-style experiment on MS MARCO v2.1 Bajaj et al. (2016). Each query has a pool of 100 retrieved passages, and the selector must compress it to $K = 5$. We compare selection baselines on the same retrieved set: Random, Top-k, MMR, Highest-α, and DPP-only. Top-k and MMR are standard baselines in modern retrieval/context-selection pipelines Ding et al. (2021). ProbDPP improves Faithfulness, F1, and ROUGE-2 over Top-k by 13.2%, 6.4%, and 16.3%, respectively, while matching Recall@K and reducing variance, including a 57.9% reduction in F1 standard deviation. Because all methods use the same retrieved candidate pool, the comparison isolates the effect of context selection under stochastic unreliability.

Adaptation to Non-Stationary Failures. Our theoretical regret analysis assumes stationary failure distributions (Section 4). To evaluate ProbDPP in dynamic environments, we conduct a time-varying experiment on HotpotQA ($N = 10$, $K = 3$) over 600 rounds. At $t = 300$, we induce an abrupt, heterogeneous shift in source reliabilities while strictly preserving the network's overall mean reliability. We evaluate 'Large' ($\Delta_\alpha = 0.30$) and "Extreme" ($\Delta_\alpha = 0.50$) mean absolute shifts. As shown in Figure 5, a fixed baseline using pre-change reliabilities retains an objective gap over the post-change evaluation period. The standard Online ProbDPP (all-history KL-UCB) responds to the changed observations but recovers slowly, as its historical observations dilute recent failures. To resolve this, we introduce an empirical extension: *Adaptive ProbDPP*, which applies KL-UCB over

Table 3: Comparison of selection methods on MS MARCO v2.1 (R2) with $K = 5$. ProbDPP improves Faithfulness, F1, and ROUGE-2 while matching Recall@K and reducing variance. Best results are bold and second-best results are underlined.

Method	Faithfulness ↑	Recall@K ↑	F1 ↑	ROUGE-2 ↑
Top-k	<u>10.88 ± 1.85</u>	**0.60**	<u>14.63 ± 1.71</u>	<u>5.35 ± 1.36</u>
MMR	10.39 ± 2.96	<u>0.57</u>	13.99 ± 2.27	4.75 ± 1.84
Random	4.44 ± 2.50	0.44	11.11 ± 1.03	2.31 ± 0.58
Highest-α	9.51 ± 1.47	0.54	13.69 ± 0.84	3.94 ± 0.53
DPP-only	2.92 ± 2.42	0.40	9.26 ± 2.15	1.45 ± 0.85
ProbDPP	**12.32 ± 1.48**	**0.60**	**15.56 ± 0.72**	**6.22 ± 0.58**

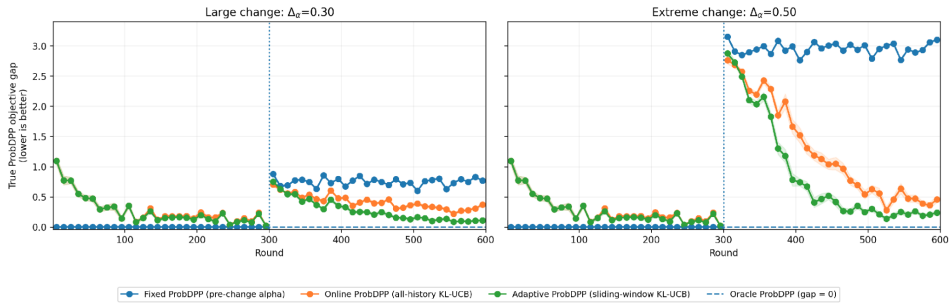


Figure 5: ProbDPP objective gap under abrupt reliability shifts at round 300, with mean reliability preserved. Panels show mean absolute shifts of 0.30 and 0.50. Sliding-window KL-UCB recovers faster than all-history KL-UCB in these experiments, while fixed pre-change reliabilities leave a persistent gap over the evaluated horizon. This experiment evaluates an empirical extension beyond the stationary regret analysis.

a sliding window of the most recent 100 outcomes. Adaptive ProbDPP rapidly forgets stale history, recovering near-oracle performance shortly after the shift. Additional downstream QA results under time-varying reliability are provided in Appendix A.9.

Query-focused summarization. To validate our claim that the ProbDPP objective handles structured dependencies (Section 3), we evaluate a strict budget scenario on QMSum (Zhong et al., 2021) with $N = 10$ simulated sources and a selection budget of $K = 3$. To test the reliability–redundancy trade-off, we assign semantically similar transcript chunks to some sources and alternative chunks to the remaining sources. We evaluate independent failures and correlated failures among the sources, preserving their marginal reliabilities. Table 4 shows that ProbDPP achieves the highest mean scores across all four metrics in both settings. Compared to the second-best method, ProbDPP improves R1/ R2 / RL/ BERT by 3.3% / 5.6% / 2.6% / 2.8% in independent failures and by 2.4% / 4.5% / 2.2% / 1.7% in correlated failures, respectively. ProbDPP selects 17.9% less redundant evidence than Highest-α, while retaining 29.5% higher average reliability than k-DPP. These results support balancing reliability and semantic coverage in this controlled summarization setting.

Synthetic reliability heterogeneity. We select $K = 3$ of $N = 10$ sources with random unit-norm features in $\mathbb{R}^3$, assigning reliabilities independently of their geometry. Across 250 seeds, we compare three reliability profiles with fixed mean $\bar{\alpha} = 0.85$ under independent Bernoulli failures. ProbDPP selects subsets with $\epsilon = 0.5$. Figure 6 shows that ProbDPP matches k-DPP under homogeneous reliability, where the reliability contribution is constant across fixed-size subsets. Under both heterogeneous profiles, ProbDPP achieves the highest mean performance among the evaluated methods. These results support reliability-aware selection even when reliability and source geometry are generated independently. There are additional results on the Appendix A.10.

Table 4: QMSum summarization under independent and correlated source failures. R1/R2/RL denote ROUGE-1/2/L; BERT denotes BERTScore. Scores are mean ± standard deviation across seeds. Higher is better; best results are bold and second-best results are underlined.

Method	R1	R2	RL	BERT
Independent failures				
ProbDPP	$\mathbf{28.43 \pm 0.27}$	$\mathbf{6.77 \pm 0.19}$	$\mathbf{17.94 \pm 0.17}$	$\mathbf{17.72 \pm 0.24}$
LexRank	$\underline{27.53 \pm 0.22}$	6.37 ± 0.21	$\underline{17.49 \pm 0.20}$	17.16 ± 0.26
k-DPP	27.15 ± 0.28	$\underline{6.41 \pm 0.18}$	17.27 ± 0.23	$\underline{17.23 \pm 0.33}$
Highest-α	26.88 ± 0.15	5.92 ± 0.10	17.32 ± 0.13	15.69 ± 0.18
MMR	26.53 ± 0.39	6.14 ± 0.22	16.91 ± 0.25	16.44 ± 0.34
Random	25.82 ± 0.34	5.68 ± 0.23	16.63 ± 0.23	15.18 ± 0.30
Lead-K	25.32 ± 0.31	5.61 ± 0.17	16.44 ± 0.19	14.40 ± 0.28
Correlated failures				
ProbDPP	$\mathbf{28.13 \pm 0.38}$	$\mathbf{6.76 \pm 0.19}$	$\mathbf{17.77 \pm 0.23}$	$\mathbf{17.56 \pm 0.34}$
LexRank	$\underline{27.47 \pm 0.34}$	6.37 ± 0.17	$\underline{17.39 \pm 0.20}$	17.05 ± 0.31
k-DPP	27.17 ± 0.39	$\underline{6.47 \pm 0.20}$	17.29 ± 0.25	$\underline{17.26 \pm 0.31}$
Highest-α	26.17 ± 0.28	5.77 ± 0.08	16.93 ± 0.15	15.28 ± 0.25
MMR	26.70 ± 0.38	6.16 ± 0.21	17.05 ± 0.25	16.55 ± 0.36
Random	25.75 ± 0.36	5.69 ± 0.17	16.62 ± 0.23	15.07 ± 0.38
Lead-K	25.20 ± 0.34	5.60 ± 0.17	16.40 ± 0.21	14.41 ± 0.36

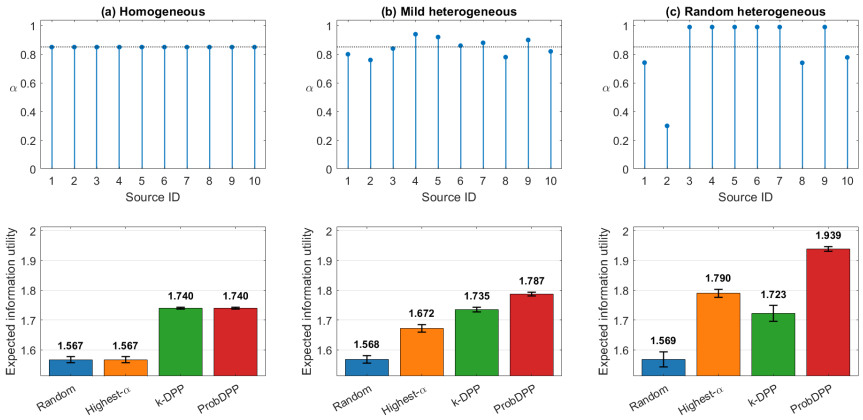


Figure 6: Effect of reliability heterogeneity at fixed mean $\bar{\alpha} = 0.85$. Top: first-seed reliability profiles; dotted lines indicate the common mean. Bottom: expected surviving-set information utility across 250 seeds, with approximately 95% confidence intervals.

6 Conclusion

We introduced ProbDPP for budgeted data selection under stochastic availability, jointly accounting for semantic diversity and source reliability. We showed that the existing DPP-based formula is ill-posed and developed a regularized objective that decomposes exactly into geometric diversity and an additive reliability reward. This decomposition depends only on marginal reliabilities and remains valid under correlated binary failures. For unknown reliabilities, we formulated a combinatorial semi-bandit problem and developed a KL-UCB-style algorithm with logarithmic regret guarantees under stationary, independent Bernoulli feedback. Experiments on question answering, retrieval-augmented generation, summarization, and visual reconstruction demonstrate the benefits of jointly considering diversity and reliability. Synthetic experiments further support these benefits when reliability is assigned independently of source geometry. Across representative benchmarks, ProbDPP improves ROUGE-2 by 16.3% in RAG, BERTScore by 13.5% in question answering, and PSNR by 3.9% in visual reconstruction over the corresponding strong baselines.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing manuscript text, formatting tables, and writing and debugging code. Generative AI was not used to develop the core research idea or the conceptual design of the proposed method. The authors reviewed all AI-assisted text and tables and verified and tested the AI-assisted code for correctness. We take responsibility for the final content of this work, including all claims, results, and artifacts produced with the assistance of generative AI.

REFERENCES

Ashkan Alinejad, Krtin Kumar, and Ali Vahdat. Evaluating the retrieval component in llm-based question answering systems. *arXiv preprint arXiv:2406.06458*, 2024.

Nima Anari, Nika Haghtalab, Seffi Naor, Sebastian Pokutta, Mohit Singh, and Alfredo Torrico. Structured robust submodular maximization: Offline and online algorithms. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3128–3137. PMLR, 2019.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Ashish Bastola, Hao Wang, Xiwen Chen, and Abolfazl Razi. Fedmil: Federated-multiple instance learning for video analysis with optimized dpp scheduling. In *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pp. 109–116. IEEE, 2024.

Xiwen Chen, Huayu Li, Peijie Qiu, Wenhui Zhu, Rahul Amin, and Abolfazl Razi. Rd-dpp: Rate-distortion theory meets determinantal point process to diversify learning data samples. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6911–6920. IEEE, 2025.

Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*, 2025.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C Park. Typos that broke the rag's back: Genetic attack on rag pipeline by simulating documents in the wild via low-level perturbations. *arXiv preprint arXiv:2404.13948*, 2024.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024.

Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. A hybrid bandit framework for diversified recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4036–4044, 2021.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. *arXiv preprint arXiv:2407.13998*, 2024.

J Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. 2006.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. Meetingbank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16409–16423, 2023.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Kurt Keutzer, and Amir Gholami. Characterizing prompt compression methods for long context inference. *arXiv preprint arXiv:2407.08892*, 2024.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*, 2023.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Tingyu Jiang, Shen Li, Yiyao Song, Lan Zhang, Hualei Zhu, Yuan Zhao, Xiaohang Xu, Kenjiro Taura, and Hao Henry Wang. Importance-aware data selection for efficient llm instruction tuning. *arXiv preprint arXiv:2511.07074*, 2025.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Alex Kulesza and Ben Taskar. Structured determinantal point processes. *Advances in neural information processing systems*, 23, 2010.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1193–1200, 2011a.

Alex Kulesza and Ben Taskar. Learning determinantal point processes. 2011b.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543. PMLR, 2015.

Sean Lamont, Christian Walder, Amir Dezfouli, Paul Montague, and Michael Norrish. 3d-prover: Diversity driven theorem proving with determinantal point processes. *arXiv preprint arXiv:2410.11133*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jiatao Li, Xinyu Hu, and Xiaojun Wan. Smart-rag: Selection using determinantal matrices for augmented retrieval. *arXiv preprint arXiv:2409.13992*, 2024.

Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Deletion-robust submodular maximization: Data summarization with "the right to be forgotten". In *International Conference on Machine Learning*, pp. 2449–2458. PMLR, 2017.

Alliot Nagle, Adway Girish, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. Fundamental limits of prompt compression: A rate-distortion framework for black-box language models. *Advances in Neural Information Processing Systems*, 37:94934–94970, 2024.

Aneta Neumann, Jakob Bossek, and Frank Neumann. Diversifying greedy sampling and evolutionary diversity optimisation for constrained monotone submodular functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 261–269, 2021.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968*, 2024.

Karthik Abinav Sankararaman. Semi-bandit feedback: A survey of results. *ARXIV CoRR*, 2016.

Ahmad Sarlak, Abolfazl Razi, Xiwen Chen, and Rahul Amin. Diversity maximized scheduling in roadside units for traffic monitoring applications. In *2023 IEEE 48th Conference on Local Computer Networks (LCN)*, pp. 1–4. IEEE, 2023.

Ahmad Sarlak, Rahul Amin, and Abolfazl Razi. Extended visibility of autonomous vehicles via optimized cooperative perception under imperfect communication. *Transportation Research Part C: Emerging Technologies*, 164:104836, 2025. doi: 10.1016/j.trc.2025.104836.

Ahmad Sarlak, Hao Wang, Rahul Amin, and Abolfazl Razi. Llm-assisted coalition formation for cooperative perception in autonomous driving. *arXiv preprint arXiv:2608.00690*, 2026.

Qiang Sun, Yuanyi Luo, Wenxiao Zhang, Sirui Li, Jichunyang Li, Kai Niu, Xiangrui Kong, and Wei Liu. Docs2kg: A human-llm collaborative approach to unified knowledge graph construction from heterogeneous documents. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 801–804, 2025.

Alex Kulesza Ben Taskar. Determinantal point processes for machine learning. *stat*, 1050:10, 2013.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Nice data selection for instruction tuning in llms with non-differentiable evaluation metric. In *Forty-second International Conference on Machine Learning*, 2025.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: budget-aware evaluation of llm reasoning strategies. *arXiv preprint arXiv:2406.06461*, 2024a.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. *arXiv preprint arXiv:2408.02103*, 2024b.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. *arXiv preprint arXiv:2408.02103*, 2024c.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122. PMLR, 2015.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025.

Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Wen-tau Yih, and Chenyan Xiong. Group-level data selection for efficient pretraining. *arXiv preprint arXiv:2502.14709*, 2025.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. A survey on data selection for llm instruction tuning. *Journal of Artificial Intelligence Research*, 83, 2025a.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025b.

Xuechen Zhang, Zijian Huang, Ege Onur Taga, Carlee Joe-Wong, Samet Oymak, and Jiasi Chen. Efficient contextual llm cascades through budget-constrained policy learning. *Advances in Neural Information Processing Systems*, 37:91691–91722, 2024.

Yuxuan Zhang, Chao Xu, Howard H Yang, Xijun Wang, and Tony QS Quek. Dpp-based client selection for federated learning with non-iid data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021.

A Appendix

A.1 Proof of Lemma 1

We show that the naïve expected log-determinant objective collapses under Bernoulli dropouts, and that regularization restores well-posedness.

Let $L \subseteq [N]$ be a selected subset and define the unregularized kernel

$$\mathbf{\Sigma}_{L,L}(\mathbf{z}_t) = M_L(\mathbf{z}_t)\, G_{L,L}\, M_L(\mathbf{z}_t), \tag{9}$$

where $z_{i,t} \sim \text{Bernoulli}(\alpha_i)$ independently and $M_L(\mathbf{z}_t) = \text{diag}(z_{i,t} : i \in L)$.

If any selected item $i \in L$ has $\alpha_i < 1$, then

$$\Pr(z_{i,t} = 0) = 1 - \alpha_i > 0, \tag{10}$$

and whenever $z_{i,t} = 0$, the diagonal matrix $M_L(\mathbf{z}_t)$ has a zero entry. Consequently, $\mathbf{\Sigma}_{L,L}(\mathbf{z}_t)$ is singular and

$$\log\det \mathbf{\Sigma}_{L,L}(\mathbf{z}_t) = -\infty \tag{11}$$

with positive probability. Therefore, these infinite terms contribute with a nonzero coefficient $(1 - \alpha_i)$ to the expected value, and we have

$$\mathbb{E}_{\mathbf{z}_t}[\log\det \mathbf{\Sigma}_{L,L}(\mathbf{z}_t)] = -\infty, \tag{12}$$

and the naïve objective is ill-posed under stochastic availability.

To resolve this issue, ProbDPP introduces the regularized kernel

$$\mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t) = (M_L(\mathbf{z}_t) + \epsilon I)\, G_{L,L}\, (M_L(\mathbf{z}_t) + \epsilon I), \qquad \epsilon > 0. \tag{13}$$

Assuming $G_{L,L} \succ 0$, Appendix A.2 shows that

$$\mathbb{E}_{\mathbf{z}_t}[\log\det \mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t)] = \log\det G_{L,L} + \sum_{i \in L} r_i(\alpha_i, \epsilon), \tag{14}$$

which is finite for all $\alpha_i \in [0, 1]$. Hence the regularized objective is well-defined and decomposes cleanly into a geometric diversity term and an additive reliability reward. □

A.2 Proof of Lemma 2

We derive a closed-form decomposition of the expected regularized diversity objective used in ProbDPP. Fix a subset $L \subseteq [N]$ and assume that the Gram submatrix $G_{L,L}$ is positive definite.[1]

The regularized kernel under stochastic availability is defined as

$$\mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t) = W_L(\mathbf{z}_t)\, G_{L,L}\, W_L(\mathbf{z}_t), \tag{15}$$

$$W_L(\mathbf{z}_t) = M_L(\mathbf{z}_t) + \epsilon I, \tag{16}$$

where $M_L(\mathbf{z}_t) = \text{diag}(z_{i,t} : i \in L)$ and $z_{i,t} \sim \text{Bernoulli}(\alpha_i)$ independently.

Since $G_{L,L} \succ 0$, it admits a Cholesky decomposition

$$G_{L,L} = R_L R_L^\top, \tag{17}$$

[1] If $G_{L,L}$ is only positive semidefinite, the same results hold after a standard ridge $G \leftarrow G + \delta I$ with $\delta > 0$.

with R_L lower triangular and invertible. Substituting into the kernel yields

$$\mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t) = W_L(\mathbf{z}_t) R_L R_L^\top W_L(\mathbf{z}_t) \tag{18}$$

$$= \big(W_L(\mathbf{z}_t) R_L\big)\big(W_L(\mathbf{z}_t) R_L\big)^\top. \tag{19}$$

Hence $\mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t)$ is positive definite for all $\mathbf{z}_t$ and $\epsilon > 0$.

Using the determinant identity $\det(AA^\top) = \det(A)^2$, we obtain

$$\log\det \mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t) = 2\log\det\big(W_L(\mathbf{z}_t) R_L\big) \tag{20}$$

$$= 2\log\det W_L(\mathbf{z}_t) + 2\log\det R_L \tag{21}$$

$$= \log\det G_{L,L} + 2\sum_{i\in L} \log(z_{i,t} + \epsilon), \tag{22}$$

where the last equality uses $\det(R_L)^2 = \det(G_{L,L})$ and the fact that $W_L(\mathbf{z}_t)$ is diagonal.

Taking expectation with respect to $\mathbf{z}_t$ gives

$$\mathbb{E}_{\mathbf{z}_t}[\log\det \mathbf{\Sigma}^{(\epsilon)}_{L,L}(\mathbf{z}_t)] = \log\det G_{L,L} + \sum_{i\in L} r_i(\alpha_i, \epsilon), \tag{23}$$

where the per-item reliability reward is

$$r_i(\alpha_i, \epsilon) \triangleq 2\Big[\alpha_i \log(1+\epsilon) + (1-\alpha_i)\log\epsilon\Big]. \tag{24}$$

This establishes the claimed decomposition. □

A.3 Ablation and analytical role of ϵ

The parameter $\epsilon > 0$ serves two roles in ProbDPP. First, it prevents the dropout-masked kernel from becoming singular. Second, under the fixed-cardinality constraint $|S| = K$, it induces an exact and interpretable trade-off between geometric diversity and source reliability.

Recall from Lemma 2 that

$$D_\epsilon(S) = \log\det(G_{S,S}) + \sum_{i\in S} 2\left[\alpha_i \log(1+\epsilon) + (1-\alpha_i)\log\epsilon\right]. \tag{25}$$

Define

$$\lambda_\epsilon \triangleq 2\log\left(\frac{1+\epsilon}{\epsilon}\right) = 2\log\left(1 + \frac{1}{\epsilon}\right). \tag{26}$$

Then

$$r_i(\alpha_i, \epsilon) = 2\log\epsilon + \lambda_\epsilon \alpha_i, \tag{27}$$

and hence, since $|S| = K$,

$$D_\epsilon(S) = \log\det(G_{S,S}) + 2K\log\epsilon + \lambda_\epsilon \sum_{i\in S} \alpha_i. \tag{28}$$

The term $2K\log\epsilon$ is independent of the selected subset. Therefore,

$$S^\star_\epsilon \in \arg\max_{|S|=K} \left\{\log\det(G_{S,S}) + \lambda_\epsilon \sum_{i\in S} \alpha_i\right\}. \tag{29}$$

Visualization of the per-item reliability term. Figure 7 visualizes the dependence of the per-item reliability term on source reliability for several values of ϵ. From equation 29, or equivalently $r_i(\alpha_i, \epsilon) = 2\log\epsilon + \lambda_\epsilon\alpha_i$, the reliability contribution is affine in α_i, with slope $\lambda_\epsilon = 2\log(1+1/\epsilon)$.

As ϵ decreases, the slope of the reliability term increases, so a given difference in source reliability produces a larger difference in the ProbDPP score. Conversely, increasing ϵ flattens the curves, reducing the influence of reliability relative to the log-determinant diversity term. Importantly, the lower absolute vertical position of the curves for small ϵ does not imply that small ϵ is unfavorable: under the fixed-cardinality constraint, the corresponding offset $2K\log\epsilon$ is identical for all feasible subsets and therefore does not affect subset selection.

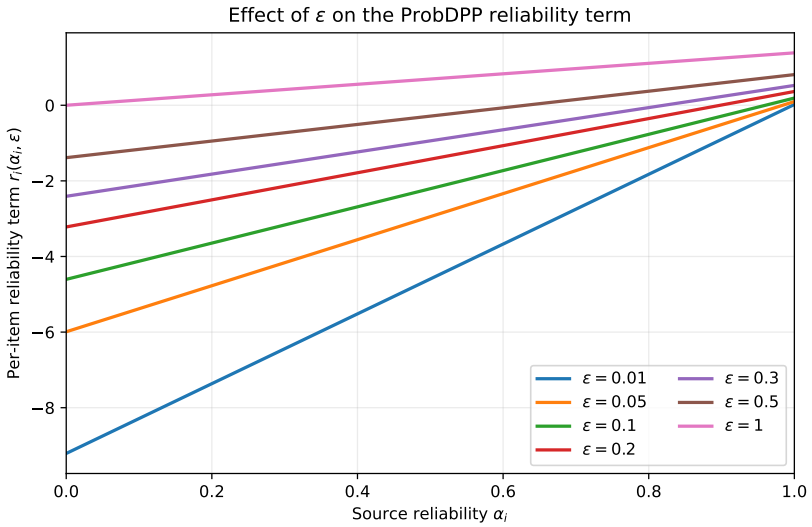


Figure 7: Per-item reliability term $r_i(\alpha_i, \epsilon)$ as a function of source reliability $\alpha_i \in [0, 1]$ for $\epsilon \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1\}$. Each curve is affine in α_i with slope $\lambda_\epsilon = 2\log(1 + 1/\epsilon)$. Smaller ϵ produces a steeper slope, increasing the relative importance of reliability differences between candidate sources, whereas larger ϵ reduces this sensitivity and shifts the selection criterion toward geometric diversity. The vertical offset $2\log\epsilon$ is constant across items for a fixed ϵ and, under the fixed-cardinality constraint, contributes only the subset-independent term $2K\log\epsilon$.

Monotonicity of the reliability weight. The effective reliability coefficient satisfies

$$\frac{d\lambda_\epsilon}{d\epsilon} = -\frac{2}{\epsilon(1+\epsilon)} < 0, \tag{30}$$

with

$$\lim_{\epsilon \to 0^+} \lambda_\epsilon = \infty, \qquad \lim_{\epsilon \to \infty} \lambda_\epsilon = 0. \tag{31}$$

Thus small ϵ strongly emphasizes source reliability, whereas large ϵ approaches diversity-only k-DPP selection.

Monotone diversity–reliability trade-off. Let $0 < \epsilon_1 < \epsilon_2$, and let S_1 and S_2 denote optimal subsets under ϵ_1 and ϵ_2, respectively. Because $\lambda_{\epsilon_1} > \lambda_{\epsilon_2}$, optimality gives

$$g(S_1) + \lambda_{\epsilon_1} A(S_1) \geq g(S_2) + \lambda_{\epsilon_1} A(S_2), \tag{32}$$
$$g(S_2) + \lambda_{\epsilon_2} A(S_2) \geq g(S_1) + \lambda_{\epsilon_2} A(S_1), \tag{33}$$

where $g(S) = \log\det(G_{S,S})$ and $A(S) = \sum_{i \in S} \alpha_i$. Adding the inequalities yields

$$(\lambda_{\epsilon_1} - \lambda_{\epsilon_2})(A(S_1) - A(S_2)) \geq 0, \tag{34}$$

and therefore

$$A(S_1) \geq A(S_2). \tag{35}$$

Substitution into the second optimality inequality further gives

$$g(S_1) \leq g(S_2). \tag{36}$$

Consequently, increasing ϵ monotonically decreases the reliability of the selected subset while monotonically increasing its log-determinant diversity.

Pairwise switching threshold. Consider two feasible subsets S and T such that $g(S) > g(T)$ but $A(S) < A(T)$, i.e., S is more diverse while T is more reliable. Define

$$\Delta_g = g(S) - g(T) > 0, \qquad \Delta_\alpha = A(T) - A(S) > 0.$$

Their score difference is

$$F_\epsilon(S) - F_\epsilon(T) = \Delta_g - \lambda_\epsilon \Delta_\alpha,$$

where

$$\lambda_\epsilon = 2 \log \left(1 + \frac{1}{\epsilon}\right).$$

Thus the two subsets have equal score when

$$\lambda_{\epsilon^\star} = \frac{\Delta_g}{\Delta_\alpha},$$

which gives

$$\epsilon^\star = \left[\exp\left(\frac{\Delta_g}{2\Delta_\alpha}\right) - 1\right]^{-1}.$$

Since λ_ϵ is strictly decreasing in ϵ,

$$\epsilon < \epsilon^\star \Rightarrow F_\epsilon(T) > F_\epsilon(S),$$

whereas

$$\epsilon > \epsilon^\star \Rightarrow F_\epsilon(S) > F_\epsilon(T).$$

Hence values below the threshold favor the more reliable subset, while values above it favor the more diverse subset.

Empirical validation. We next verify both the downstream effect of ϵ and the diversity–reliability mechanism predicted by the preceding analysis. We use HotpotQA with $N = 10$ candidate passages and select $K = 3$ passages by exhaustive search over all $\binom{10}{3} = 120$ feasible subsets. We evaluate heterogeneous source-reliability profiles with mean reliability $\bar{\alpha} \in \{0.5, 0.6, 0.7, 0.8\}$. For each reliability setting, we sweep ϵ over several orders of magnitude while keeping the questions, embeddings, downstream model, and Bernoulli failure realizations fixed across ϵ. Results are averaged over 10 random seeds and 120 questions.

Effect on downstream QA performance. Figure 8 shows that an intermediate ϵ gives the strongest downstream performance across the evaluated reliability profiles. Peak Token-F1 occurs at approximately $\epsilon = 0.3$–0.7, with similar performance at neighboring intermediate values. Small ϵ strongly favors reliability, whereas large ϵ weakens the reliability contribution and approaches diversity-only selection. The results therefore support an intermediate operating range rather than a single universally optimal ϵ.

Verification of the diversity–reliability mechanism. Figure 9 agrees with the analytical trade-off: increasing ϵ raises the selected log-determinant diversity while lowering average selected reliability. The best Token-F1 operating points lie between the extremes, where diversity has increased without the larger reliability losses observed at high ϵ. Further increases in geometric diversity therefore need not improve downstream QA performance.

A.4 Algorithmic details

Algorithm 1 summarizes the online implementation of ProbDPP when the source reliabilities are unknown. At each round, the learner constructs KL-UCB confidence bounds for the Bernoulli reliability parameters, maps these optimistic reliability estimates through the reward function $r_i(\cdot, \epsilon)$, and selects the subset that maximizes the sum of the log-determinant diversity term and the optimistic reliability reward. After observing semi-bandit feedback from the selected sources, the empirical reliability estimates are updated and used in subsequent rounds.

A.5 Proof of Theorem 1

For each item $i \in [N]$, define the item-wise gaps

$$\Delta_i^{\min} \triangleq \min\{f(A^\star) - f(a) : a \in \mathcal{A},\ a_i = 1,\ f(a) < f(A^\star)\},$$

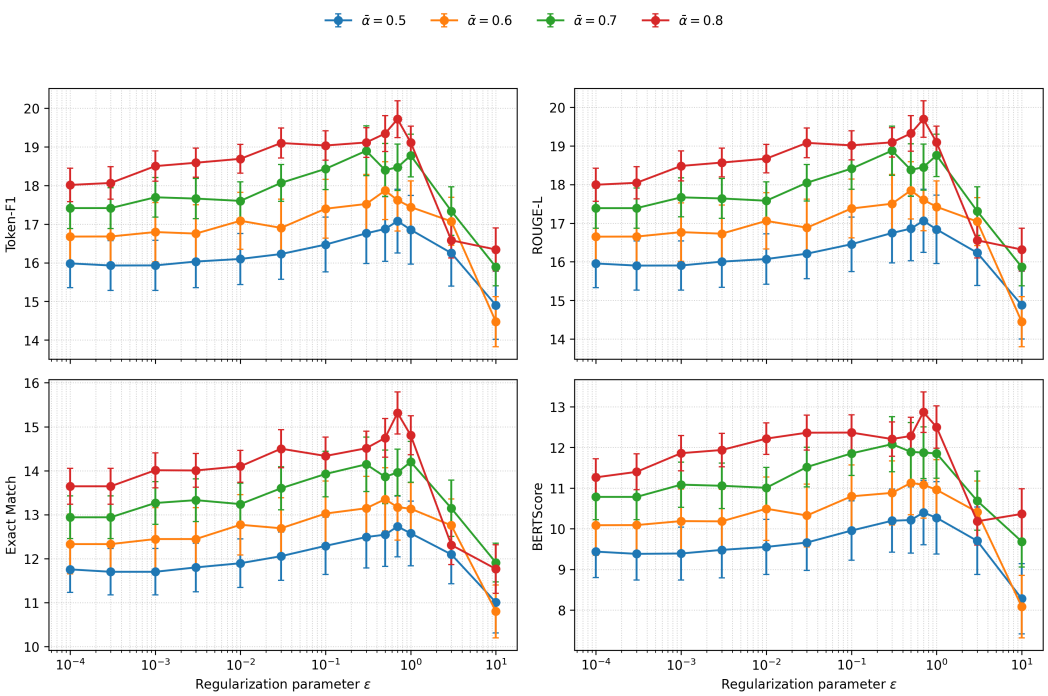


Figure 8: Each curve corresponds to a heterogeneous source-reliability profile with mean $\bar{\alpha} \in \{0.5, 0.6, 0.7, 0.8\}$. Points show the mean across random seeds and error bars show the corresponding standard deviation. Very small ϵ places excessive weight on source reliability, whereas very large ϵ suppresses the reliability contribution and approaches diversity-only selection. Across Token-F1, ROUGE-L, Exact Match, and BERTScore, the strongest downstream performance occurs in an intermediate range of ϵ, demonstrating that neither extreme of the reliability–diversity trade-off is optimal for downstream QA.

with $\Delta_i^{\min} = +\infty$ if no strictly suboptimal action contains i, and

$$\Delta_i^{\max} \triangleq \max\{f(A^\star) - f(a) : a \in \mathcal{A},\ a_i = 1,\ f(a) < f(A^\star)\},$$

with $\Delta_i^{\max} = 0$ if no suboptimal action contains i.

The per-round reward is

$$f(a) = g(a) + \langle a, \theta \rangle, \tag{37}$$

where $\theta_i = r_i(\alpha_i, \epsilon)$ and $\alpha_i \in [0, 1]$ is unknown. Let $A^\star \in \arg\max_{a \in \mathcal{A}} f(a)$ and define the Reg_T

$$Reg_T = \sum_{t=1}^{T} \big(f(A^\star) - f(A_t)\big). \tag{38}$$

From Appendix A.2,

$$r_i(\alpha, \epsilon) = 2 \log \epsilon + \beta_\epsilon \alpha, \tag{39}$$

$$\beta_\epsilon \triangleq 2 \log\left(\tfrac{1+\epsilon}{\epsilon}\right) > 0, \tag{40}$$

so $r_i(\cdot, \epsilon)$ is strictly increasing.

Let $U_i^\alpha(t)$ be the KL-UCB index for α_i and define

$$\tilde{\theta}_i(t) = r_i(U_i^\alpha(t), \epsilon). \tag{41}$$

Standard KL-UCB analysis ensures that with high probability,

$$\theta_i \le \tilde{\theta}_i(t) \quad \forall i, t. \tag{42}$$

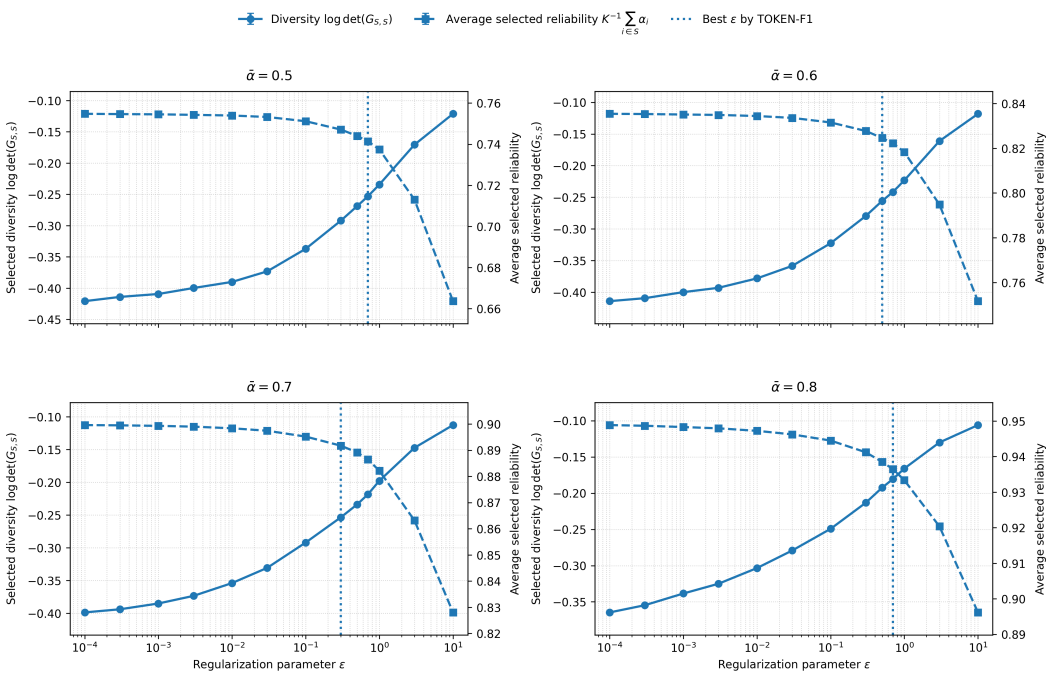


Figure 9: The solid curve reports the selected geometric diversity $\log\det(G_{S,S})$ on the left axis, and the dashed curve reports the average true reliability of the selected subset, $K^{-1}\sum_{i\in S}\alpha_i$, on the right axis. Each panel corresponds to a heterogeneous reliability profile with the displayed mean $\bar{\alpha}$. The vertical dotted line marks the value of ϵ attaining the highest Token-F1 in Figure 8. As predicted analytically, increasing ϵ monotonically increases the selected log-determinant diversity while decreasing the average selected reliability. The best downstream operating points occur in the transition region where substantial diversity has been gained without yet incurring the large reliability loss observed at high ϵ.

Let $\Delta(a) = f(A^\star) - f(a)$. If A_t is suboptimal, then

$$\Delta(A_t) \le \sum_{i:A_{t,i}=1} \big(\tilde{\theta}_i(t) - \theta_i\big). \tag{43}$$

Hence there exists at least one selected item i such that

$$\tilde{\theta}_i(t-1) - \theta_i \ge \Delta_i^{\min}/K, \tag{44}$$

where $\Delta_i^{\min}$ is the minimal gap of actions containing item i.

Define the bad optimism indicator

$$B_{i,t} = \mathbb{I}\Big\{A_{t,i} = 1,\ \tilde{\theta}_i(t) - \theta_i \ge \Delta_i^{\min}/K\Big\}. \tag{45}$$

Then

$$\mathbb{I}\{A_t \ne A^\star\} \le \sum_{i=1}^{N} B_{i,t}. \tag{46}$$

By equation 40, the event $\tilde{\theta}_i(t) - \theta_i \ge \Delta_i^{\min}/K$ is equivalent to

$$U_i^\alpha(t-1) \ge \alpha_i^+ \triangleq \min\Big\{1,\ \alpha_i + \tfrac{\Delta_i^{\min}}{K\beta_\epsilon}\Big\}. \tag{47}$$

KL-UCB counting arguments yield

$$\mathbb{E}\left[\sum_{t=1}^{T} B_{i,t}\right] \le \frac{\log T + c\log\log T}{\mathrm{kl}(\alpha_i\|\alpha_i^+)} + O(1). \tag{48}$$

Algorithm 1 ProbDPP

Input: horizon T
Action set: $\mathcal{A} = \{a \in \{0,1\}^N : \|a\|_1 = K\}$
Initialize $n_i(0) = 0$, $\hat{\alpha}_i(0) = 0$ for all $i \in \{1, \dots, N\}$
for $t = 1$ **to** T **do**
  for $i = 1$ **to** N **do**
    if $n_i(t-1) \geq 1$ **then**
      Compute $U_i^{\alpha}(t)$:

$$U_i^{\alpha}(t) = \max_{u \in [0,1]} u$$

$$\text{s.t.} \quad n_i(t-1)\,\mathrm{kl}\big(\hat{\alpha}_i(t-1) \,\|\, u\big) \leq log t + c \log\log t$$

    else
      Set $U_i^{\alpha}(t) = 1$
    end if
    Set $\tilde{r}_i(t) = r_i\big(U_i^{\alpha}(t), \epsilon\big)$
  end for
  Select A_t:

$$A_t \in \arg\max_{a \in \mathcal{A}} \{\, g(a) + \langle a, \tilde{r}(t) \rangle \,\}$$

  Observe $\{z_{i,t} : A_{t,i} = 1\}$
  for each i with $A_{t,i} = 1$ **do**
    Update $n_i(t)$ and $\hat{\alpha}_i(t)$
  end for
end for

with $\Delta_i^{\max} = 0$ if no suboptimal action contains item i, D_{KL} denotes Bernoulli KL divergence and $c > 0$ is the KL-UCB exploration constant. Whenever $B_{i,t} = 1$, the selected action contains item i, and therefore

$$f(A^\star) - f(A_t) \leq \Delta_i^{\max}.$$

Using equation 46, we obtain the refined charging bound

$$\mathrm{Reg}_T \leq \sum_{i=1}^{N} \Delta_i^{\max} \sum_{t=1}^{T} B_{i,t}.$$

Taking expectations and applying the KL-UCB counting bound gives

$$\mathbb{E}[\mathrm{Reg}_T] \leq \sum_{i=1}^{N} \Delta_i^{\max} \frac{\log T + c \log\log T}{\mathrm{kl}(\alpha_i \| \alpha_i^+)} + O(1).$$

A.6 Proof of Canonical Lower Bound in 4

We prove the lower bound by reducing a special case of the reliability-aware ProbDPP problem to a classical Bernoulli bandit.

Restricted instance. Consider the special case $K = 1$ and choose the Gram matrix G such that $g(a)$ is constant for all singleton sets $S = \{i\}$. Then $g(a)$ does not affect regret differences, and each action corresponds to selecting a single item $i \in [N]$. The per-round objective reduces to

$$f(i) = r_i(\alpha_i, \epsilon), \tag{49}$$

and the problem becomes an N-armed stochastic bandit with Bernoulli feedback $z_{i,t} \sim \mathrm{Bernoulli}(\alpha_i)$ and mean reward $\mu_i \triangleq r_i(\alpha_i, \epsilon)$.

Let $i^\star = \arg\max_i \mu_i$ and define the reward gaps

$$\Delta_i \triangleq \mu_{i^\star} - \mu_i > 0, \qquad i \neq i^\star. \tag{50}$$

Since $r_i(\alpha, \epsilon)$ is affine and increasing in α, these gaps are equivalently proportional to reliability gaps $\alpha_{i^\star} - \alpha_i$.

Information-theoretic argument. Fix any policy π and a suboptimal arm $i \neq i^\star$. Construct an alternative instance α' that differs from α only in coordinate i, with $\alpha'_i > \alpha_{i^\star}$, so that arm i is uniquely optimal. Let P_α and $P_{\alpha'}$ denote the distributions over the interaction history induced by π under α and α'. Since the two instances differ only on arm i,

$$D_{\mathrm{KL}}(P_\alpha \| P_{\alpha'}) = \mathbb{E}_\alpha[n_i(T)]\, \mathrm{kl}(\alpha_i \| \alpha'_i), \tag{51}$$

where $n_i(T)$ is the number of times arm i is selected up to time T.

Consequence for regret. Standard change-of-measure arguments imply that any policy achieving subpolynomial regret must satisfy

$$\mathbb{E}_\alpha[n_i(T)] \;\gtrsim\; \log T \quad \text{for all } i \neq i^\star, \tag{52}$$

and therefore

$$\mathbb{E}_\alpha[\mathrm{Reg}_T] \;\gtrsim\; \sum_{i \neq i^\star} \Delta_i \log T \;\geq\; (N-1)\, \delta_{\min}\, \log T, \tag{53}$$

where $\delta_{\min} \triangleq \min_{i \neq i^\star} \Delta_i$.

Instance-dependent form. Equivalently, by the Bernoulli Lai–Robbins lower bound,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\alpha[\mathrm{Reg}_T]}{\log T} \geq \sum_{i \neq i^\star} \frac{\Delta_i}{\mathrm{kl}(\alpha_i \| \alpha_{i^\star})}. \tag{54}$$

Implication for ProbDPP. Since this Bernoulli bandit is a special case of the reliability-aware ProbDPP semi-bandit formulation, no algorithm can achieve $o(\log T)$ regret uniformly over instances. Combined with Theorem 1, this shows that ProbDPP achieves order-optimal regret in T, up to constants and gap-dependent factors.

A.7 Stationary Online Reliability Learning

We complement the regret analysis in Section 4 with a controlled stationary experiment designed to match its feedback assumptions. The purpose of this experiment is not to model a particular downstream dataset, but to isolate the online reliability-learning component of ProbDPP under the setting covered by the theoretical analysis.

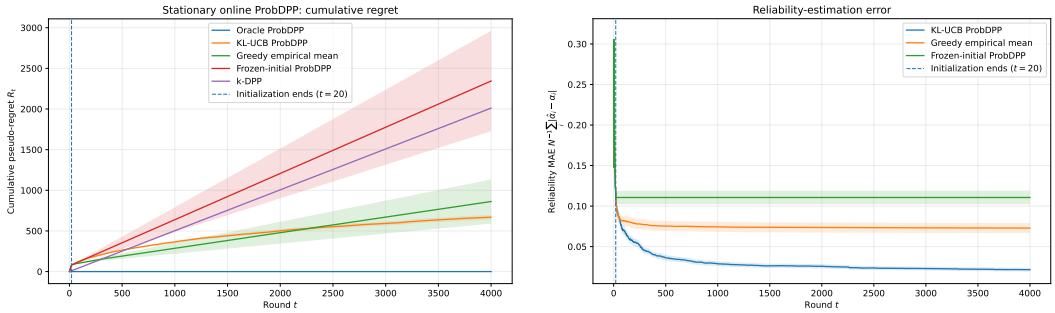
We consider $N = 10$ persistent sources and select $K = 3$ sources at each round for a horizon of $T = 4000$ rounds. Source reliabilities are fixed throughout the experiment and feedback is generated independently as $z_{i,t} \sim \mathrm{Bernoulli}(\alpha_i)$. We use $\epsilon = 0.1$ and a single fixed Gram matrix G throughout all rounds, methods, and random seeds. The Gram matrix is constructed from clustered source embeddings so that the instance contains a nontrivial trade-off between geometric diversity and source reliability. Results are averaged over 50 random seeds. Within each seed, all methods are evaluated using the same potential Bernoulli outcomes.

We compare five policies. *Oracle ProbDPP* selects using the true reliabilities α_i and therefore serves as the zero-regret reference. *KL-UCB ProbDPP* is the online method proposed in Section 3 and selects using optimistic reliability estimates. *Greedy empirical mean* replaces the KL-UCB indices with the current empirical reliability estimates and therefore performs no explicit optimism-driven exploration. *Frozen-initial ProbDPP* estimates reliability during initialization and then keeps these estimates fixed for all subsequent rounds. Finally, *k-DPP* ignores reliability and selects using only the geometric log-determinant term.

To give each online estimator an initial observation of every source, the first 20 rounds use a balanced initialization schedule. After initialization, each learning method observes feedback only from the sources that it selects. Thus, an estimator can improve the estimate of source i only when that source is selected, making the comparison between KL-UCB and greedy empirical-mean selection an explicit test of the exploration–exploitation mechanism.

Figure 10 reports cumulative pseudo-regret and reliability-estimation error. We use the same pseudo-regret definition as in Section 4,

$$R_t = \sum_{\tau=1}^{t} \left[f(S^\star) - f(S_\tau) \right],$$



(a) Cumulative pseudo-regret.

(b) Reliability-estimation MAE.

Figure 10: Controlled stationary online-learning experiment with $N = 10$, $K = 3$, $T = 4000$, $\epsilon = 0.1$, and 50 random seeds. The dashed vertical line marks the end of the balanced initialization period at $t = 20$. Solid curves show the mean across seeds and shaded regions show approximate 95% confidence intervals. **(a)** KL-UCB ProbDPP incurs an initial exploration cost but subsequently accumulates substantially less pseudo-regret than Frozen-initial ProbDPP and diversity-only k-DPP, and its long-horizon regret is lower than that of greedy empirical-mean selection. **(b)** KL-UCB continues reducing reliability-estimation error as feedback is accumulated, whereas the greedy and frozen alternatives plateau at larger errors.

Table 5: Stationary online reliability learning with $N = 10$, $K = 3$, $T = 4000$, and 50 seeds. Results are mean $\pm$ standard deviation across seeds.

Method	R_T	Final MAE	Final ℓ_∞	Non-oracle rate	Avg. oracle gap
Oracle ProbDPP	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
KL-UCB ProbDPP	668.649 ± 99.592	0.022 ± 0.008	0.094 ± 0.055	0.358 ± 0.076	0.167 ± 0.025
Greedy empirical mean	861.931 ± 977.281	0.073 ± 0.022	0.279 ± 0.103	0.369 ± 0.213	0.215 ± 0.244
Frozen-initial ProbDPP	2344.338 ± 2224.125	0.111 ± 0.030	0.319 ± 0.089	0.441 ± 0.206	0.586 ± 0.556
k-DPP	2011.110 ± 0.000	N/A	N/A	0.667 ± 0.000	0.503 ± 0.000

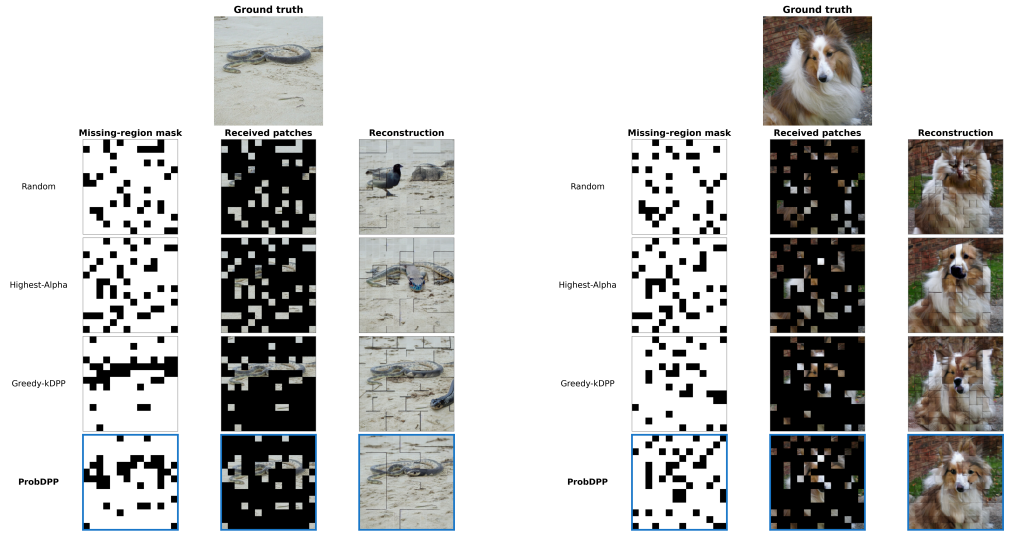
where $S^\star$ is an optimal size-K subset under the true reliabilities. Reliability-estimation accuracy is measured by

$$\mathrm{MAE}_\alpha(t) = \frac{1}{N}\sum_{i=1}^{N} |\widehat{\alpha}_i(t) - \alpha_i| \,.$$

The two panels expose the mechanism underlying the online method. KL-UCB initially explores sources whose reliability is uncertain, which can increase short-horizon regret relative to a purely greedy policy. Continued exploration, however, provides observations of sources that a greedy policy may cease selecting. Consequently, KL-UCB progressively improves its reliability estimates, while the empirical-mean and frozen estimators can retain persistent estimation error. This difference is reflected in the cumulative objective: the policies that do not maintain explicit exploration accumulate a larger long-run oracle gap.

For completeness, Table 5 reports the final estimation errors and action-level statistics. In addition to MAE, we report $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_\infty$, which measures the largest remaining source-level reliability error. We also report the non-oracle selection rate and average per-round oracle gap. The non-oracle rate should be interpreted as an action-level diagnostic rather than as a direct decomposition of regret into pure exploration and exploitation costs.

These results provide an empirical complement to the stationary regret analysis: under the fixed independent Bernoulli model assumed by the theory, explicit optimism enables ProbDPP to continue learning source reliabilities from semi-bandit feedback and to reduce the objective loss relative to policies that either freeze or greedily exploit their reliability estimates. We emphasize that this controlled experiment validates the online-learning mechanism under the stationary model; it is distinct from the time-varying HotpotQA experiment in Section 5 and Appendix A.9, which evaluates an empirical nonstationary extension beyond the stationary regret guarantee.



(a) (b)

Figure 11: Qualitative examples of BLIP-assisted Stable Diffusion v1.5 inpainting under stochastic patch loss. The ground-truth image is shown above each example. Rows correspond to selection methods; columns show the missing-region mask, received patches, and reconstructed image. White mask regions denote missing pixels. The displayed configuration is $\bar{\alpha} = 0.8$, $K = 49$, and $\epsilon = 0.3$.

A.8 Additional Robust Visual Transmission and Reconstruction

While Section 5 demonstrates the effectiveness of ProbDPP for visual token selection using Masked Autoencoders (MAE), we further evaluate its versatility across advanced generative inpainting and reconstruction models. Table 6 summarizes the quantitative performance of ProbDPP compared to baselines using BLIP-assisted Stable Diffusion v1.5 (BLIP-SD1.5), Mask-Aware Transformer (MAT), and Large Mask Inpainting (LaMa). Consistent with our MAE findings, maximizing delivery alone (Highest-α) or diversity alone (Greedy-kDPP) yields sub-optimal reconstruction quality. By jointly optimizing for both semantic coverage and transmission reliability, ProbDPP achieves the highest mean PSNR and SSIM among the evaluated selectors for each of the three reconstruction backends in Table 6. Relative to the strongest competing selector for each metric, the PSNR improvements are approximately 0.094, 0.211, and 0.272 dB for BLIP-SD1.5, MAT, and LaMa, respectively; the corresponding absolute SSIM improvements are 0.0025, 0.0084, and 0.0175. These results provide evidence that the selection benefit extends beyond the MAE reconstruction pipeline, although its magnitude depends on the reconstruction backend.

Table 6: Quantitative evaluation of robust visual transmission and reconstruction under stochastic access links. ProbDPP consistently achieves the highest PSNR and SSIM across BLIP-SD1.5, MAT, and LaMa architectures compared to diversity-only (Greedy-kDPP), reliability-only (Highest-Alpha), and random selection baselines.

Method	**BLIP-SD1.5**		**MAT**		**LaMa**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Random	13.312 ± 0.033	0.3731 ± 0.0015	13.107 ± 0.012	0.3181 ± 0.0004	17.901 ± 0.024	0.5528 ± 0.0005
Highest-Alpha	13.698 ± 0.138	0.3930 ± 0.0031	13.276 ± 0.207	0.3319 ± 0.0051	18.282 ± 0.144	0.5720 ± 0.0032
Greedy-kDPP	13.337 ± 0.045	0.3773 ± 0.0019	13.206 ± 0.015	0.3284 ± 0.0008	19.751 ± 0.035	0.5787 ± 0.0007
ProbDPP	$\mathbf{13.792 \pm 0.025}$	$\mathbf{0.3955 \pm 0.0009}$	$\mathbf{13.487 \pm 0.050}$	$\mathbf{0.3403 \pm 0.0014}$	$\mathbf{20.023 \pm 0.045}$	$\mathbf{0.5962 \pm 0.0011}$

A.9 Additional Results under Non-Stationary Failures

We complement the objective-gap analysis with downstream QA results on HotpotQA under time-varying source reliability. We compare Fixed ProbDPP, which retains pre-change reliability estimates; Online ProbDPP, which updates estimates using all-history KL-UCB; and Adaptive Prob-

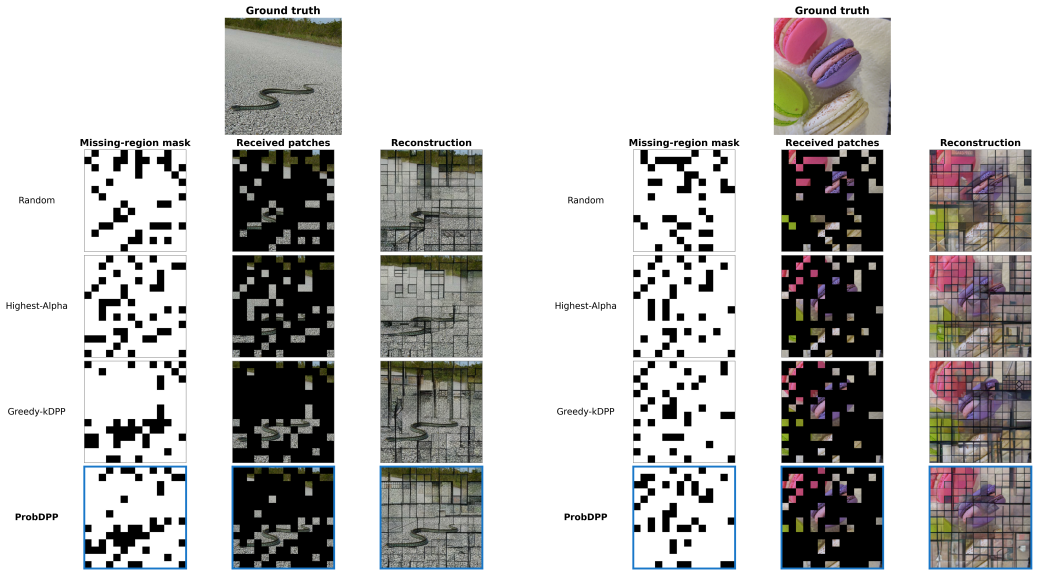


Figure 12: Qualitative examples of MAT inpainting under stochastic patch loss. The ground-truth image is shown above each example. Rows correspond to selection methods; columns show the missing-region mask, received patches, and reconstructed image. The displayed configuration is $\bar{\alpha} = 0.8$, $K = 49$, and $\epsilon = 0.3$.

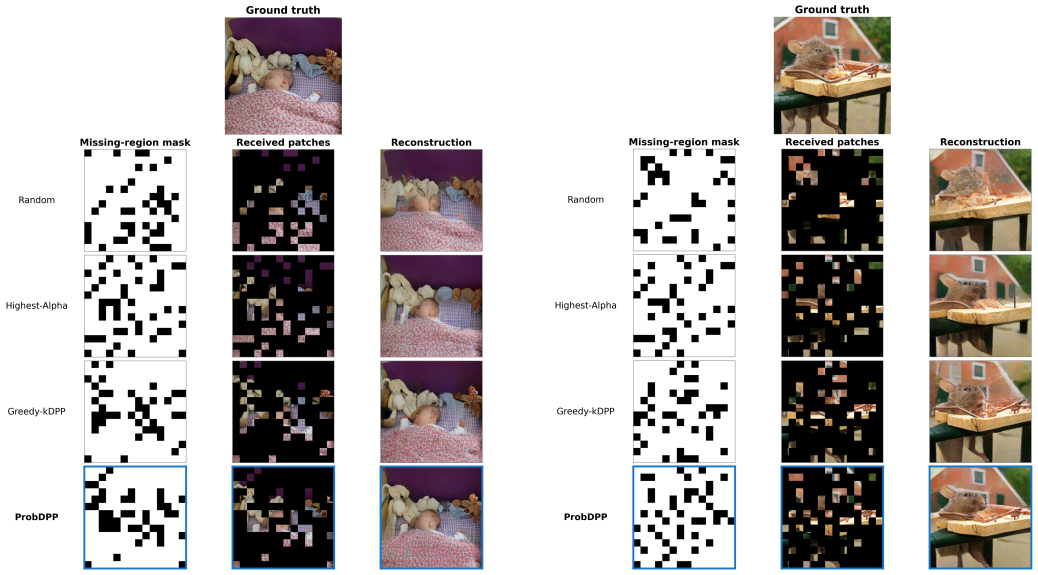


Figure 13: Qualitative examples of LaMa inpainting under stochastic patch loss. The ground-truth image is shown above each example. Rows correspond to selection methods; columns show the missing-region mask, received patches, and reconstructed image. The displayed configuration is $\bar{\alpha} = 0.8$, $K = 49$, and $\epsilon = 0.3$.

DPP, which uses sliding-window KL-UCB. Oracle ProbDPP uses the current true reliabilities and serves as a reference. We also include diversity-only k-DPP, Highest-α, and Random selection.

Results. Table 7 shows that updating reliability estimates improves downstream QA performance relative to keeping them fixed. Online ProbDPP increases Token-F1 from 16.26 to 17.14, while Adaptive ProbDPP further improves it to 17.45. Adaptive ProbDPP achieves the highest reported

Table 7: Downstream QA performance on HotpotQA under time-varying source reliability. Higher is better for all metrics. Bold denotes the highest reported mean among non-oracle methods; Oracle ProbDPP is shown separately as a reference.

Method	Token-F1	ROUGE-L	BERTScore	EM
k-DPP	14.14 ± 0.80	14.14 ± 0.80	8.01 ± 0.88	10.84 ± 0.82
Highest-α	15.15 ± 0.81	15.15 ± 0.81	8.82 ± 0.92	11.09 ± 0.91
Random	16.13 ± 1.47	16.11 ± 1.46	9.97 ± 1.54	–
Fixed ProbDPP	16.26 ± 0.75	16.26 ± 0.75	10.02 ± 0.73	12.29 ± 0.75
Online ProbDPP	17.14 ± 1.11	17.14 ± 1.11	10.63 ± 1.09	13.80 ± 0.98
Adaptive ProbDPP	$\mathbf{17.45 \pm 1.08}$	$\mathbf{17.45 \pm 1.08}$	$\mathbf{11.17 \pm 0.89}$	$\mathbf{14.01 \pm 1.08}$
Oracle ProbDPP	18.70 ± 0.64	18.70 ± 0.64	12.47 ± 0.73	14.29 ± 0.60

mean scores among non-oracle methods across all four metrics, with relative Token-F1 gains of 7.3% over Fixed ProbDPP and 1.8% over Online ProbDPP. These results are consistent with the benefit of discounting stale feedback, although the additional gains over all-history KL-UCB are modest and statistical significance has not been established. A gap to Oracle ProbDPP remains, particularly in Token-F1 and BERTScore. This experiment evaluates an empirical nonstationary extension and is not covered by the stationary regret guarantees.

A.10 Additional experimental results

Robustness to Correlated and Heterogeneous Failures. We evaluate ProbDPP ($N = 10, K = 3$) on HotpotQA using 20 random seeds. We assign highly heterogeneous marginal reliabilities ($\alpha_i \in [0.05, 0.99]$, mean 0.70). We evaluate two regimes: (1) independent Bernoulli failures, and (2) correlated failures, where sources are partitioned into two groups of five. Each group shares a latent uniform variable $u_{\text{group}} \sim \text{Uniform}(0, 1)$, and a source survives if $u_{\text{group}} < \alpha_i$. This simulates common-cause system outages (e.g., a shared edge cache going offline) while strictly preserving the true marginal survival probability $\Pr(z_i = 1) = \alpha_i$. ProbDPP achieves the highest mean scores across all four metrics in both regimes (Table 8). Under independent failures, its Token-F1 is 18.79, compared with 17.43 for Highest-α and 16.09 for k-DPP. Under correlated failures, its Token-F1 remains similar at 18.70. These results demonstrate empirical robustness to the tested common-cause failure model; the marginal-only dependence established for the ProbDPP objective does not imply that downstream QA performance is invariant to correlation.

Table 8: Comparison of selection methods under independent and correlated failure modes with highly heterogeneous source reliabilities on HotpotQA. ProbDPP successfully maintains superior performance even when failures are strongly correlated.

Failure Mode	**Method**	**Token-F1**	**ROUGE-L**	**BERTScore**	**EM**
Independent	Random	16.70 ± 0.88	16.69 ± 0.88	10.24 ± 0.96	12.78 ± 0.69
	Highest-α	17.43 ± 0.43	17.41 ± 0.43	10.68 ± 0.39	13.02 ± 0.41
	k-DPP	16.09 ± 0.58	16.09 ± 0.58	9.77 ± 0.66	11.78 ± 0.50
	ProbDPP	$\mathbf{18.79 \pm 0.60}$	$\mathbf{18.77 \pm 0.60}$	$\mathbf{12.17 \pm 0.61}$	$\mathbf{13.89 \pm 0.54}$
Correlated	Random	17.31 ± 0.85	17.29 ± 0.85	10.80 ± 0.92	13.13 ± 0.65
	Highest-α	17.51 ± 0.38	17.49 ± 0.38	10.66 ± 0.33	13.07 ± 0.37
	k-DPP	16.33 ± 0.54	16.33 ± 0.54	9.95 ± 0.52	11.78 ± 0.55
	ProbDPP	$\mathbf{18.70 \pm 0.33}$	$\mathbf{18.68 \pm 0.33}$	$\mathbf{11.99 \pm 0.32}$	$\mathbf{13.70 \pm 0.34}$

Multi-Source Selection under Stochastic Availability.

Additionally, we evaluate ProbDPP in multi-source selection settings with stochastic source availability using the HotpotQA distractor benchmark Yang et al. (2018), using llm "llama3" and standard QA metrics. As shown in Table 9, ProbDPP consistently outperforms all baselines on token-overlap,

Table 9: Multi-source selection under stochastic source availability on HotpotQA (distractor setting, $N = 10$, $K = 3$). Higher is better; best in bold.

	ProbDPP	Highest-α	k-DPP	Random
Token-F1	**34.14**	22.12	31.20	26.86 ± 3.58
ROUGE-L	**34.14**	22.47	31.34	26.87 ± 3.61
BERTScore	**11.27**	8.4	7.67	4.82 ± 4.32
EM	**25.00**	17.00	24.00	21.00 ± 3.60

semantic similarity, and exact-answer metrics. In the $K = 3$ setting, ProbDPP achieves Token-F1 and ROUGE-L scores of 34.14, substantially exceeding both the reliability-Highest-α baseline (22.12 / 22.47) and the diversity-only (k-DPP) baseline (31.20 / 31.34). The gains are substantial for semantic and factual accuracy: ProbDPP attains a BERTScore of 11.27 and an exact match of 25, compared to 8.4–7.67 and 17–24 for baselines, and also exhibits lower variance than random selection baseline. figs. 14a to 14d further demonstrate robustness to stochastic availability as α varies. Each query is associated with $N = 10$ candidate sources, from which the model selects a subset of size $K = 5$, while each source independently survives with probability α, modeling realistic partial failures. The results are averaged over 10 runs and 120 scenarios. Across all metrics, ProbDPP achieves the best or most stable performance, with Exact Match steadily improving from 17 ± 5.4 at $\alpha = 0.5$ to 22.5 ± 7.2 at $\alpha = 0.8$, while reliability-only and diversity-only baselines remain flat or degrade. Similar trends hold for ROUGE and BERTScore, indicating improved semantic alignment and preservation of relevant evidence in partial failures. The sensitivity analysis in the selection budget $K \in \{3, 4, 5, 6\}$ (figs. 15a to 15d) shows that ProbDPP consistently outperforms or matches the baselines across all budgets, with particularly strong advantages in low-budget regimes. For example, at $K = 4$, ProbDPP achieves EM of 21 ± 5.2 and ROUGE-L of 25.06 ± 6.74, outperforming both Highest-α and k-DPP, which suffer from lower accuracy and higher variance. These results indicate that ProbDPP's gains are not driven by larger context sizes, but by its reliability-aware diversity objective, which selects subsets that are both informative and likely to remain available. Overall, the results confirm that jointly modeling reliability and diversity yields robust and accurate multi-source selection under constrained and stochastic conditions.

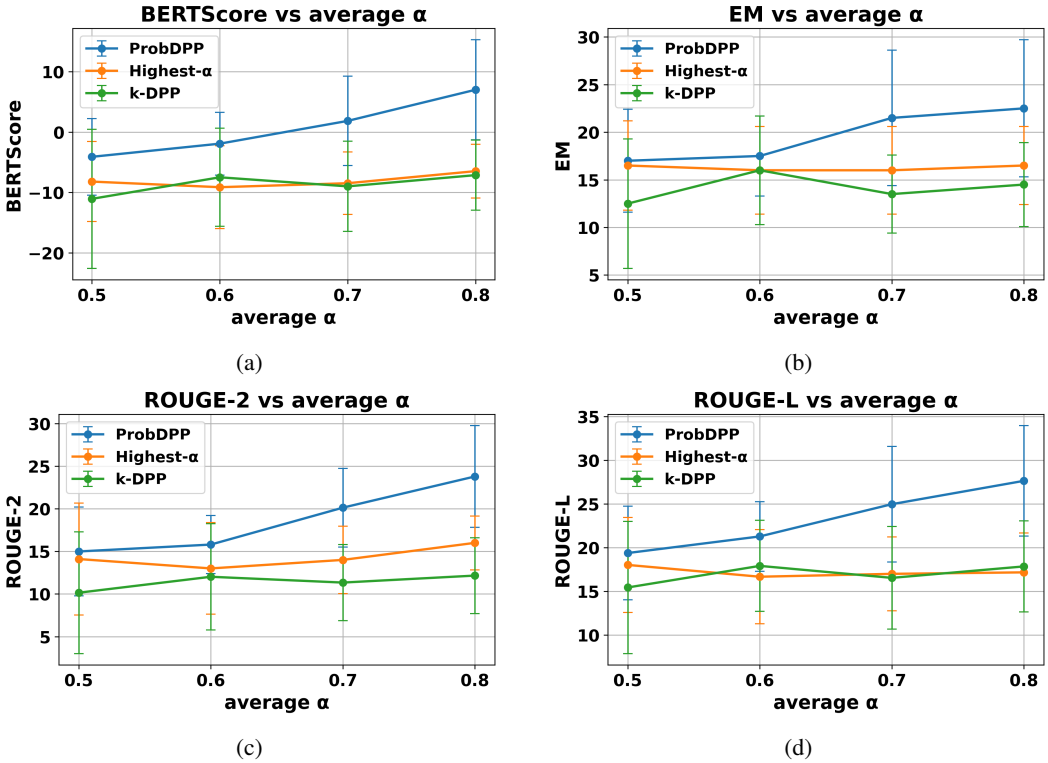


Figure 14: Robustness to stochastic source availability. Performance vs. survival average probability α with $N = 10$ and $K = 5$. Error bars show mean $\pm$ std.

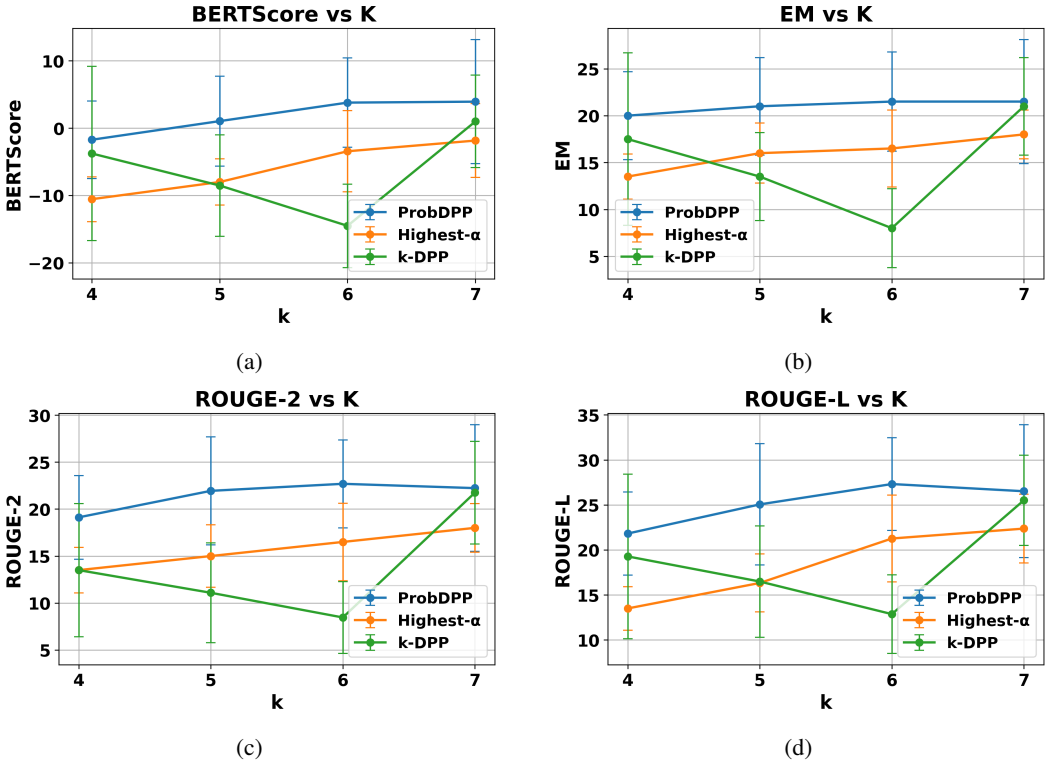


Figure 15: Effect of selection budget under stochastic availability. Performance vs. selection budget K with $N = 10$ candidate sources. Error bars show mean $\pm$ std.