AVERT: AN ACTION-WISE ELIGIBILITY FUNNEL FOR OFFLINE CLARIFICATION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline evaluation can promote a clarification policy when its estimated value is driven by actions for which the log contains little effective support. We present Advance-or-refuse Validation under Effective support for Rare Typed actions (AVERT), an action-wise eligibility funnel for a frozen family of clarification policies. AVERT screens every typed clarification action using effective sample size, candidate–logger propensity alignment, and an unclipped importance-weight-tail diagnostic before applying clipped doubly robust ranking. The method separates selection from execution: selection either advances the lowest-loss eligible policy or returns a typed refusal that identifies the action and failed support condition; execution then applies an advanced fixed policy as a one-step value-of-clarification controller. We formalize this interface and give a finite-family conditional concentration result for the clipped ranking objective. The resulting object makes missing action-specific evidence an explicit decision outcome rather than an implicit property of an offline winner.

1 INTRODUCTION

Clarification policies make a consequential allocation decision: they may answer immediately, request a particular kind of information, abstain, or defer. An offline policy selector must make a different decision first: whether available logged evidence supports advancing a candidate policy at all. This distinction is especially important when a candidate's ranking depends on a rare typed action. A favorable scalar estimate can obscure that the action responsible for the estimate has weak overlap with the logging policy.

AVERT makes that precondition explicit. Before ranking, it creates an eligibility record for every candidate and every typed clarification action. A candidate advances only when its action-level diagnostics meet the fixed screening rule. A failed diagnostic excludes that candidate; when no candidate remains, AVERT returns a typed refusal of the form $\textsc{refuse}(a, \text{condition})$. Ranking is consequently performed inside an eligible family rather than used as evidence that the family itself is supported. Passing this screen is an admission rule, not a guarantee of overlap, unbiased evaluation, safety, or prospective performance.

This framing differs from baseline-relative high-confidence improvement and support-constrained offline reinforcement learning, which quantify or constrain improvement relative to a baseline (Thomas et al., 2015; Laroche et al., 2019). It also differs from uncertainty estimates attached to an individual prediction (Malinin & Gales, 2018; Kuhn et al., 2023): AVERT's output is a decision about whether a finite policy family is eligible for prospective use. Its action-specific record preserves the source of a refusal, which distinguishes insufficient support for one clarification type from a global failure of value estimation.

The paper contributes (i) an action-wise eligibility predicate that precedes clipped doubly robust ranking, (ii) an explicit two-stage execution semantics that separates offline advancement from online clarification control, and (iii) a finite-family result for the clipped ranking objective after a fixed support screen. Together these components make support a visible control variable in offline clarification-policy selection.

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.

2 CLARIFICATION DECISION PROBLEM

Let a clarification controller operate on a finite-horizon partially observable Markov decision process

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, \mathcal{L}, \gamma).$$

We use a factored latent state $s = (s^{\text{int}}, s^{\text{haz}}, s^{\text{coop}})$ for user intent, domain-specific hazards, and response cooperativeness. The action set is

$$\mathcal{A} = \{\texttt{answer}, \texttt{abstain}, \texttt{defer}\} \cup \{\texttt{ask}_k\}_{k=1}^{K},$$

where $\{\texttt{ask}_k\}_{k=1}^{K}$ is a fixed inventory of typed clarification actions. The inventory is chosen before policy selection; each type can have a domain-specific wording template, while the decision variable remains the type.

For harm channels $\mathcal{H}$ with nonnegative severity weights w_h, define the terminal-action loss

$$\mathcal{L}(a, s) = \sum_{h \in \mathcal{H}} w_h p_h(s, a) + c(a), \tag{1}$$

where $p_h(s, a)$ is the channel-h harm probability and $c(a)$ is the cost of the terminal action. Given a belief b_t, write

$$\bar{\mathcal{L}}(a, b_t) = \mathbb{E}_{s \sim b_t}[\mathcal{L}(a, s)] \quad \text{and} \quad \bar{\mathcal{L}}^*(b_t) = \min_{a \in \{\texttt{answer},\texttt{abstain},\texttt{defer}\}} \bar{\mathcal{L}}(a, b_t).$$

3 AVERT

3.1 ONE-STEP VALUE OF CLARIFICATION

For a clarification type k, the value of clarification is the expected decrease in optimal terminal loss after observing a response, less the interaction cost:

$$\text{VoC}(k, b_t) = \mathbb{E}_{o' \sim P(\cdot \mid b_t, \texttt{ask}_k)} \left[\bar{\mathcal{L}}^*(b_t) - \bar{\mathcal{L}}^*(b_t^{o',k}) \right] - c_k. \tag{2}$$

Here $b_t^{o',k}$ is the posterior belief after response o' to question type k. The action-value inputs are

$$J(\texttt{ask}_k, b_t) = \bar{\mathcal{L}}^*(b_t) - \text{VoC}(k, b_t)$$

alongside the terminal-action costs. AVERT does not prescribe a response-model architecture or policy parameterization. Rather, each candidate $\pi \in \Pi$ is a frozen stochastic action rule $\pi(a \mid x_t, b_t, J_t)$, where J_t collects these action-value inputs; a deterministic candidate is a point-mass special case. The same frozen rule is evaluated on logged contexts and, after advancement, applied during execution. Thus a candidate may use the supplied action values to select an action, but AVERT neither refits it nor substitutes a new controller after ranking.

3.2 CLIPPED DOUBLY ROBUST RANKING

The advancement problem uses logged one-step contexts

$$\mathcal{D} = \{(x_i, a_i, \ell_i, \mu(\cdot \mid x_i))\}_{i=1}^{n},$$

where x_i is a decision context containing the frozen-policy inputs needed to evaluate π, a_i is the logged action, ℓ_i is its loss label, and $\mu(\cdot \mid x_i)$ is the recorded logger-propensity vector. Write $\mu_i = \mu(a_i \mid x_i)$ for the propensity of the logged action. The vector supplies $\mu(a \mid x_i)$ for each action used by the action-wise screen. For a candidate policy π, let

$$w_i(\pi) = \frac{\pi(a_i \mid x_i)}{\mu_i}, \qquad \widetilde{w}_i(\pi) = \min\{w_i(\pi), \tau_{\text{clip}}\}.$$

With a loss model $\widehat{Q}$, AVERT ranks candidates using the clipped doubly robust objective (Dudík et al., 2014; Jiang & Li, 2016):

$$\widehat{V}_{\text{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{Q}(x_i, \pi) + \widetilde{w}_i(\pi) \big(\ell_i - \widehat{Q}(x_i, a_i) \big) \right], \tag{3}$$

where $\widehat{Q}(x_i, \pi) = \sum_a \pi(a \mid x_i) \widehat{Q}(x_i, a)$. Clipping bounds the contribution of extreme corrections. The clipped objective is the target of the ranking result stated below.

3.3 ACTION-WISE ELIGIBILITY

The support screen is evaluated on a screening split fixed separately from the ranking split. For each clarification action a, define the raw action-specific weights

$$w_{i,a}(\pi) = \mathbb{1}\{a_i = a\}\frac{\pi(a \mid x_i)}{\mu(a \mid x_i)}$$

and effective sample size

$$\mathrm{ESS}_a(\pi) = \frac{\left(\sum_i w_{i,a}(\pi)\right)^2}{\sum_i w_{i,a}^2(\pi)}, \tag{4}$$

with value zero when the denominator is zero. Let $\mathrm{Align}_a(\pi, \mu)$ be the Pearson correlation, on the screening split, between $\log\max\{\pi(a \mid x_i), \varepsilon\}$ and $\log\max\{\mu(a \mid x_i), \varepsilon\}$ for a fixed numerical floor $\varepsilon > 0$. If either vector has zero empirical variance, set this diagnostic to $-\infty$. Let $T_q^{\mathrm{raw}}(w_a)$ be the empirical q-quantile of nonzero raw action-specific weights, for a fixed $q \in (0, 1)$, and set it to $+\infty$ when no such weight is present. These conventions make a degenerate action unsupported rather than leaving its eligibility undefined.

For thresholds (δ, ρ, τ) fixed before ranking, the admissible family is

$$\begin{aligned}\Pi_{\mathrm{ad}}(\delta, \rho, \tau) = \big\{\pi \in \Pi :\ & \mathrm{ESS}_a(\pi) \geq \delta, \\ & \mathrm{Align}_a(\pi, \mu) \geq \rho, \\ & T_q^{\mathrm{raw}}(w_a) \leq \tau, \quad \forall a \in \mathcal{A}_{\mathrm{ask}}\big\}.\end{aligned} \tag{5}$$

The tail diagnostic is computed before clipping, so it diagnoses entry into the clipping region rather than restating the clipped estimator. AVERT returns ADVANCE$(\widehat{\pi})$ with

$$\widehat{\pi} \in \arg\min_{\pi \in \Pi_{\mathrm{ad}}} \widehat{V}_{\mathrm{DR}}(\pi)$$

when the admissible set is nonempty; otherwise it returns a refusal with the failed action-level conditions.

The three diagnostics define a fixed eligibility predicate only. In particular, correlation of candidate and logger propensities is not an overlap proof, and a raw-weight quantile does not bound observations above that quantile. Eligibility therefore does not establish an unbiased estimate, a safety improvement, or a prospective-performance guarantee; it records the screen that was required before the clipped ranking comparison.

3.4 EXECUTION SEMANTICS

AVERT has two noninterchangeable stages. The selection stage receives a frozen candidate family, a support-screening split, a ranking split with recorded propensities, a loss model, and screening thresholds. It outputs either an advanced fixed policy or a typed refusal; it never changes the candidate family, logger, or thresholds while ranking. The execution stage is entered only for an advanced policy and applies that same $\widehat{\pi}$; it neither re-ranks candidates nor replaces the advanced action rule. At turn t, it receives the current belief b_t, a response model $r(\cdot \mid x_{0:t}, \mathtt{ask}_k, b_t)$, a specified number of response samples, and terminal-action costs. It constructs J_t by approximating Eq. (2), applies $\widehat{\pi}(\cdot \mid x_t, b_t, J_t)$, observes a response if the selected action asks, updates the belief, and repeats until a terminal action or the turn limit.

Algorithm 1 AVERT selection and execution interface.

Require: Candidate family Π, screening split, ranking split, thresholds (δ, ρ, τ)
Require: For an advanced policy $\widehat{\pi}$: response model r, response-sample count m, terminal-action costs, and turn limit H
Compute action-wise eligibility diagnostics for every $\pi \in \Pi$
$\Pi_{\mathrm{ad}} \leftarrow$ candidates satisfying Eq. (5)
if $\Pi_{\mathrm{ad}} = \varnothing$ **then**

    return REFUSE(failed action-condition pairs)
end if
$\widehat{\pi} \leftarrow \arg\min_{\pi \in \Pi_{\mathrm{ad}}} \widehat{V}_{\mathrm{DR}}(\pi)$
Advance the fixed policy $\widehat{\pi}$
for each execution turn $t \le H$ **do**
    Construct J_t by estimating Eq. (2) with m samples from $r(\cdot \mid x_{0:t}, \mathtt{ask}_k, b_t)$ for each typed question
    Sample (or take the deterministic action) $a_t \sim \widehat{\pi}(\cdot \mid x_t, b_t, J_t)$; execute a_t and update the belief after a clarification response
end for

4 Conditional Ranking Analysis

The following statement concerns the clipped objective used by the selector. It isolates the sampling contribution once the candidate family and support screen are fixed.

Proposition 1. *Condition on the screening split and its resulting fixed finite nonempty admissible family Π_{ad}. Suppose the ranking split is independent of the screening split, its observations are independent (not necessarily identically distributed), the loss model $\widehat{Q}$ is fixed independently of the ranking split, $\ell_i, \widehat{Q}(x_i, a) \in [0, L]$, and $\widetilde{w}_i(\pi) \in [0, \tau_{\mathrm{clip}}]$ for every i and $\pi \in \Pi_{\mathrm{ad}}$. Let $Z_i(\pi)$ denote the ith summand in Eq.* (3) *and define $V_{\mathrm{clip}}(\pi) = n^{-1} \sum_i \mathbb{E}[Z_i(\pi)]$. Let $\widehat{\pi}$ minimize the empirical clipped objective. Then, with conditional probability at least $1 - \alpha$,*

$$V_{\mathrm{clip}}(\widehat{\pi}) - \min_{\pi \in \Pi_{\mathrm{ad}}} V_{\mathrm{clip}}(\pi) \le 2(1 + 2\tau_{\mathrm{clip}}) L \sqrt{\frac{\log(2|\Pi_{\mathrm{ad}}|/\alpha)}{2n}}.$$

Proof. Conditional on the screening split, for a fixed policy each summand in Eq. (3) lies in $[-\tau_{\mathrm{clip}} L, (1 + \tau_{\mathrm{clip}})L]$ and therefore has range at most $(1 + 2\tau_{\mathrm{clip}})L$. Hoeffding's inequality for independent bounded observations and a union bound give a uniform deviation bound of the displayed square-root term for all policies in Π_{ad}. Comparing the empirical minimizer with a population minimizer contributes the deviation once for each policy, yielding the factor two. □

The proposition formalizes the selection guarantee for the clipped ranking functional. The eligibility predicate complements it by making action-level support and raw-tail diagnostics explicit before a policy is eligible to compete.

5 Related Work

Clarification and information-gathering decisions are naturally expressed through partially observable control and value-of-information formulations (Kaelbling et al., 1998; Howard, 1966; Raiffa & Schlaifer, 1961). Doubly robust estimators provide a standard route for estimating policy value from logged interactions (Dudík et al., 2014; Jiang & Li, 2016). High-confidence improvement and SPIBB-style constraints use baseline-relative controls (Thomas et al., 2015; Laroche et al., 2019); AVERT instead exposes action-wise support as an eligibility condition for a finite family before scalar ranking. Selective prediction likewise makes rejection explicit (Chow, 1970; Geifman & El-Yaniv, 2017), whereas AVERT returns a typed refusal at the policy-selection boundary.

6 Conclusion

AVERT formalizes offline advancement as an eligibility decision rather than a direct consequence of a favorable value estimate. Its action-wise support predicate, typed refusal, and separated execution semantics give a clear interface for frozen clarification-policy families. The conditional finite-family analysis characterizes the clipped ranking objective after the support screen, making the role of action-specific evidence explicit in offline selection.

AI Use Statement

Generative AI was used to assist drafting, editing, formatting, and revision of the formal-method and mathematical exposition in this manuscript. The manuscript reports no generated synthetic data, experiments, result analysis, or implementation. The authors reviewed the AI-assisted text and remain responsible for the final content.

References

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. doi: 10.1214/14-STS500. URL https://doi.org/10.1214/14-STS500.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Ronald A Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074. URL https://doi.org/10.1109/TSSC.1966.300074.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, 1961. URL https://katalog.bibliothek.kit.edu/bib/532299.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

A Evaluation Contract

An empirical evaluation of AVERT fixes the typed action inventory, the candidate family, the support-screening split, the ranking split, and all screening thresholds before policy ranking. The evaluation record identifies the logging propensities, the unit of resampling for any uncertainty interval, the denominator for every binary rate, and the aggregation rule for every reported outcome. When a candidate is refused, the record retains the action and support condition that produced the refusal. These fields make the selection-stage decision independently inspectable from the behavior of a policy that has already advanced.