How Much Schema Can a Reader Use? Capacity-Adaptive Test-Time Exposure for Out-of-Domain Multi-Intent Detection

Anonymous authors
Paper under double-blind review

Abstract

When a frozen language model, or reader, predicts intents from label descriptions, showing more labels can help it find the correct ones but also introduce distractions. How many labels to show depends on both the query and the reader. We introduce Capacity-Adaptive Pool Exposure (CAPE) for multi-intent detection in new domains, without labeled target examples or access to reader internals. A source-trained selector estimates whether a candidate pool contains all required labels. Tests built from target label descriptions measure how well each reader handles additional labels, and prediction agreement under reordered descriptions refines this estimate for each query. Across 72 settings spanning nine local readers and four domains, CAPE improves macro exact-set accuracy (Exact) by 12.2% relative to fixed top-20 exposure, displaying 11.3 labels on average for the final prediction. With matched calibration and per-query call budgets, it improves Exact by 2.0% overall and 2.2% on held-out readers relative to reader-calibrated coverage thresholding. These results show that measuring reader responses to additional labels improves exposure decisions beyond estimating label coverage alone.

1 Introduction

A new intent domain often provides label names and descriptions before labeled utterances. Description prompting lets a frozen language model, which we call a reader, predict an intent set from a ranked subset of intent descriptions (Parikh et al., 2023; Hong et al., 2024), but usually fixes its depth across queries and readers.

Schema exposure involves a trade-off between covering gold intents and introducing distracting alternatives. Figure 1 shows that the best tested depth varies across readers and schemas, without increasing consistently with model size. Coverage estimates alone do not describe how additional labels affect the reader.

We introduce Capacity-Adaptive Pool Exposure (CAPE) to adapt exposure to both the query and the reader. A source-trained selector predicts gold coverage, schema-derived tests estimate a reader-specific capacity distribution, and agreement across schema orders refines this distribution for each query. A coverage–retention score combines these estimates to select the final exposure.

Across 72 local settings, CAPE improves macro Exact by 2.73 percentage points over three-view top-20 exposure with equal per-query call budgets, and by 0.86 points over reader-calibrated coverage thresholding with equal calibration and query-call budgets. Gains over full-schema exposure are largest among lower-performing readers. On the two larger intent schemas, all nine readers show positive Exact differences over both the matched threshold policy and a simplified CAPE variant.

Our contributions are threefold: a coverage–preservation formulation of exposure selection; label-free reader calibration with query-level adaptation; and validation of its predictive and decision value on held-out readers.

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.

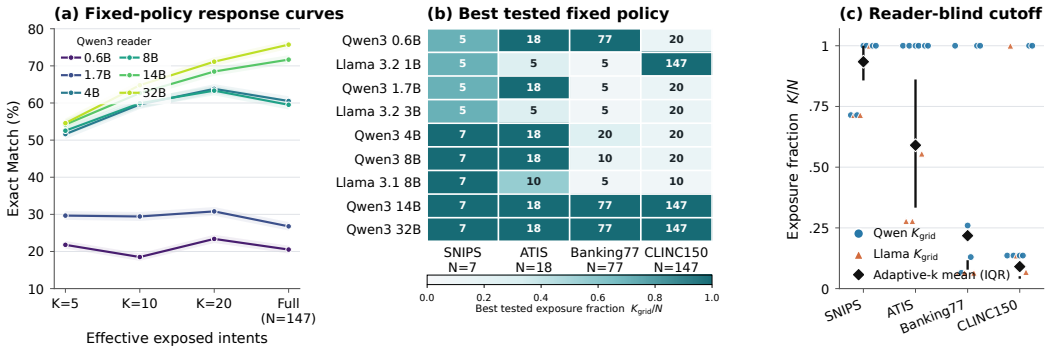


Figure 1: **Reader–schema variation in useful exposure.** (a) Qwen3 fixed-policy Exact on CLINC150 with 95% bootstrap confidence intervals (CIs). (b) Best tested effective depth among $\{5, 10, 20, N\}$; color shows its schema fraction. (c) Reader-blind Adaptive-k versus reader-dependent preferred exposure. Results average Blend/Mix; capped duplicate depths collapse to Full and ties favor smaller pools. This finite-grid diagnostic is not an exhaustive oracle and is not used by CAPE.

2 Related Work

Intent classification. Description prompting supports unseen schemas (Hu et al., 2024), while pretrained intent encoders improve transfer (Sung et al., 2023). Multi-intent models emphasize joint intent–slot or graph reasoning (Qin et al., 2020; Song et al., 2022); Divide–Solve–Combine Prompting (DSCP) decomposes the query, solves its component requests, and combines the resulting intent predictions (Qin et al., 2025). These methods do not calibrate exposure to a newly deployed reader.

Adaptive retrieval. Retrieval-augmented generation (RAG) and its adaptive variants select evidence for generation (Lewis et al., 2020; Asai et al., 2024; Jeong et al., 2024; Jiang et al., 2023). Adaptive-k uses similarity gaps (Taguchi et al., 2025); Learning Distraction-Aware Retrieval (LDAR) trains a retriever using reader outcomes (Shim et al., 2026). Here, retrieved items define the allowed output labels. Omitting a gold intent prevents exact prediction, while distracting alternatives can cause false positives. CAPE calibrates unseen readers using only their outputs, without fitting reader-specific source policies.

Behavioral measurement and adaptation. TinyBenchmarks estimates aggregate scores from selected benchmark items (Maia Polo et al., 2024); our schema-derived interventions instead estimate retention to choose exposure. Related test-time and prompt adaptation methods use unlabeled evidence with varying access assumptions (Wang et al., 2021; Boudiaf et al., 2022; Xiao et al., 2025; Fang et al., 2026; Shu et al., 2022). CAPE updates a capacity distribution outside the reader while keeping the reader weights and classification instructions fixed.

3 Problem Setup and Diagnosis

Let $\mathcal{I}_{\mathrm{T}} = \{i_1, \ldots, i_N\}$ be a held-out target schema with N intents. Each intent has a name and natural-language descriptions. Target utterances are excluded from training and model selection. At deployment, a query x expresses an unknown gold set $Y \subseteq \mathcal{I}_{\mathrm{T}}$, with $1 \leq |Y| \leq 3$.

A fixed sentence encoder ranks the intents by cosine similarity to x. Let $\pi_x(j)$ be the index of the intent at rank j. For an exposure depth $K \in \{1, \ldots, N\}$, the reader receives the first K ranked intents:

$$\mathcal{I}_K(x) = \{i_{\pi_x(1)}, \ldots, i_{\pi_x(K)}\}. \tag{1}$$

The frozen reader m returns a prediction $\widehat{Y}_K(x, m) \subseteq \mathcal{I}_K(x)$.

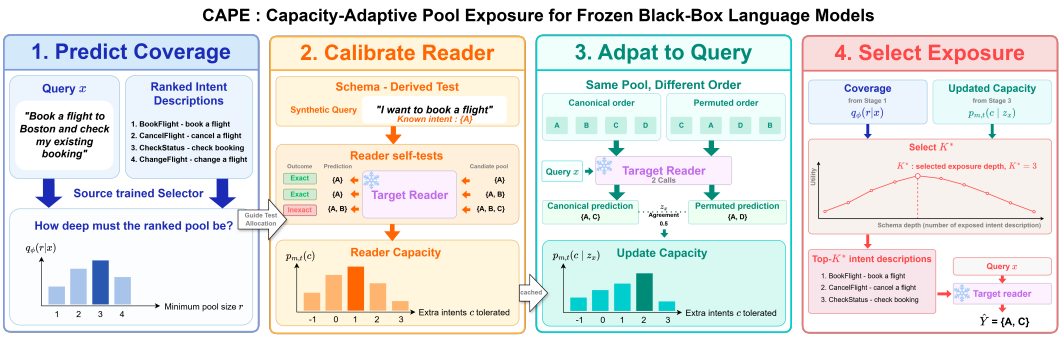


Figure 2: **CAPE overview.** (1) Predict the minimum ranked-pool depth needed to cover all gold intents. (2) Aggregate schema-derived self-tests into reader capacity, with unlabeled-batch coverage posteriors guiding test allocation. (3) Use prediction agreement under two orders of the same pool to obtain the updated capacity $p_{m,t}(c \mid z_x)$. (4) Combine coverage and updated capacity in a decision surrogate to select $K^\star$ and make one final prediction. Examples and distributions are schematic; $c = -1$ denotes failure at first gold coverage.

Let $\mathrm{rank}_{\pi_x}(y)$ denote the position of gold intent y in this ranking. The minimum gold-cover rank $R(x, Y)$ is the smallest depth that contains the complete gold set:

$$R(x, Y) = \min\{K : Y \subseteq \mathcal{I}_K(x)\} = \max_{y \in Y} \mathrm{rank}_{\pi_x}(y). \tag{2}$$

If $R > K$, exact prediction is impossible because a gold intent is unavailable. If $R \leq K$, the reader can still make an error because the displayed alternatives affect its decision (Zheng et al., 2024).

Let $[\cdot]$ denote the indicator of a statement. Define $A_K = [R \leq K]$ as complete gold coverage and $E_K = [\widehat{Y}_K = Y]$ as exact prediction, suppressing x and m where clear. Because the reader can output only displayed intents, $E_K = 1$ requires $A_K = 1$. Thus,

$$\Pr(E_K = 1 \mid x, m) = \underbrace{\Pr(A_K = 1 \mid x)}_{\text{gold coverage}} \underbrace{\Pr(E_K = 1 \mid A_K = 1, x, m)}_{\text{reader preservation}}. \tag{3}$$

This factorization requires no independence assumption. As K increases, gold coverage cannot decrease, whereas correctness after coverage can. Exposure selection must therefore balance these effects for each query and reader.

Comparing fixed exposure depths. Figure 1 compares $\{5, 10, 20, N_t\}$ after collapsing capped duplicates. Across 36 reader–schema pairs, the Exact-maximizing depth $K^*_{\mathrm{grid}}(m, t)$ varies within each schema. Ties favor smaller pools; this diagnostic is not used by CAPE.

4 CAPE

Figure 2 summarizes the four stages of CAPE. The final exposure decision combines the coverage posterior $q_\phi(r \mid x)$ from Stage 1 with the query-updated capacity distribution $p_{m,t}(c \mid z_x)$ from Stage 3.

4.1 Source Coverage Posterior

A frozen 768-dimensional `all-mpnet-base-v2` encoder (Reimers & Gurevych, 2019; Song et al., 2020) embeds queries and intent descriptions. A two-layer Transformer processes the cosine-ranked schema using rank positions, similarity scores, and adjacent score gaps, as detailed in Appendix B.

Let ϕ denote the selector parameters and $q_\phi(r \mid x, \mathcal{I}_S)$ its predicted probability that the minimum gold-cover rank is r under source schema $\mathcal{I}_S$. Training minimizes the negative log probability of the observed gold-cover rank, averaged over labeled source examples (x, Y):

$$\mathcal{L}_{\mathrm{cover}}(\phi) = -\mathbb{E}_{(x,Y)} \log q_\phi(R(x, Y) \mid x, \mathcal{I}_S). \tag{4}$$

We omit the schema argument when it is clear. Let $C_\phi(K \mid x)$ be the predicted probability of complete gold coverage at depth K, and let $K_0(x)$ be the most likely gold-cover rank:

$$C_\phi(K \mid x) = \sum_{r=1}^{K} q_\phi(r \mid x), \qquad K_0(x) = \arg\max_{r \in \{1,\dots,N\}} q_\phi(r \mid x). \tag{5}$$

4.2 Label-Free Reader Preservation from Schema-Derived Tests

The coverage posterior estimates the depth needed to include the gold set. We next estimate the reader's tolerance for additional intents. A source-trained response model selects the initial self-tests; the deployed reader's own outcomes then provide the evidence for its capacity distribution.

For two evaluated prefix depths $L < U$, the model estimates the probability of an exact prediction at U given an exact prediction at L. It learns a difficulty $d_j(x)$ for each added rank j and a scalar tolerance θ_m for each source reader m. With response-model parameters ψ and logistic sigmoid σ, this probability is modeled as

$$G_{\theta_m}(U \mid L, x) := \Pr_{\psi}(E_U = 1 \mid E_L = 1, x, m) = \prod_{j=L+1}^{U} \sigma(\theta_m - d_j(x)). \tag{6}$$

At a fixed reader tolerance, a larger difficulty reduces predicted retention. The parameters are fitted to endpoint outcomes using the conditional composite likelihood (Varin et al., 2011) in Appendix B.3. The product models correctness at the upper endpoint U, allowing an intermediate prediction to be incorrect. For the deployed reader, we construct synthetic queries from target descriptions, each with one to three known intents. For each probe b, we fix its gold set and evaluate selected prefix depths at or above its minimum gold-cover rank R_b. We approximate these responses using a boundary B_b: predictions are modeled as exact from R_b to B_b and inexact at larger depths. The state $B_b = R_b - 1$ represents failure when the gold set first becomes available.

Let $\mathcal{D}_b$ contain the queried depths and observed outcomes. These observations narrow the possible boundary to an interval $[\ell_b, u_b]$. We retain this uncertainty by assigning equal probability to every remaining integer boundary v:

$$\Pr(B_b = v \mid \mathcal{D}_b) = \frac{[\ell_b \le v \le u_b]}{u_b - \ell_b + 1}. \tag{7}$$

Calibration uses coverage posteriors from an unlabeled deployment batch $\mathcal{U}_t$, the complete target test split in headline experiments. By default, each reader–schema pair receives 64 self-test evaluation calls. The source response model selects initial probes; remaining calls maximize expected improvement in exposure decisions averaged over $\mathcal{U}_t$. Allocation requires neither annotations nor reader predictions on these queries.

For each probe, $B_b - R_b$ measures how many additional intents the reader tolerates after gold coverage. We aggregate these estimates into a distribution over capacity c, with $c = -1$ denoting failure at first coverage. For reader m and schema t, the survival function $S_{m,t}(d)$ gives the estimated probability of tolerating at least $d \ge 0$ additional intents.

We first average $\Pr(B_b - R_b \ge d \mid \mathcal{D}_b)$ over probes for which $R_b + d \le N_t$. Weighted isotonic regression (De Leeuw et al., 2010) then makes this survival curve non-increasing. We obtain the capacity distribution $p_{m,t}(c)$ on $\{-1, 0, \dots, N_t - 1\}$ from differences between adjacent survival values, as detailed in Appendix C.2. Its mean $\hat{c}_{m,t}$ summarizes reader capacity:

$$\widehat{c}_{m,t} = \mathbb{E}_{c \sim p_{m,t}}[c]. \tag{8}$$

This mean is used only for diagnostics; exposure selection uses the full distribution.

For a gold-cover rank $r \le K$, the pool contains $K - r$ additional intents beyond gold coverage. We weight retention at this excess depth by the predicted rank probability and sum over covered ranks:

$$\widetilde{U}_{m,t}(K \mid x) = \sum_{r=1}^{K} q_\phi(r \mid x) S_{m,t}(K - r). \tag{9}$$

We use this coverage–retention score as a decision surrogate, without equating it to exact-prediction probability. Section 6 evaluates its predictive and decision value.

Let K_{reader} be the depth selected before query adaptation:

$$K_{\text{reader}}(x, m, t) = \arg \max_{K \in \{K_0(x), \ldots, N_t\}} \widetilde{U}_{m,t}(K \mid x). \tag{10}$$

The predicted anchor $K_0(x)$ sets the minimum allowed depth.

4.3 Query-Level Generalized Bayes Update

The capacity distribution summarizes synthetic tests. To adapt it to each query, we check whether predictions remain stable when the same pool is reordered. We select K_p by expected value of information (EVI) (Rainforth et al., 2024), the expected improvement in the final exposure decision; Appendix C.3 gives details. The reader predicts from $\mathcal{I}_{K_p}(x)$ in canonical and deterministically permuted orders, yielding $\widehat{Y}^{\text{can}}$ and $\widehat{Y}^{\text{perm}}$. Their Dice agreement z_x is

$$z_x = \frac{2|\widehat{Y}^{\text{can}} \cap \widehat{Y}^{\text{perm}}|}{|\widehat{Y}^{\text{can}}| + |\widehat{Y}^{\text{perm}}|} \in [0, 1]. \tag{11}$$

Dice (Dice, 1945) is symmetric, accounts for different set sizes, and preserves partial agreement. Let $s_c(x, K_p)$ model order stability at depth K_p under capacity state c. This working model predicts stability when the pool covers the gold set and its excess depth remains within capacity c. When coverage is incomplete, it assigns probability $1/2$. No additional parameters are trained.

Let $p_{m,t}(c \mid z_x)$ denote the query-updated capacity distribution. Using the observed agreement z_x as a soft Bernoulli observation gives

$$p_{m,t}(c \mid z_x) \propto p_{m,t}(c)\, s_c(x, K_p)^{z_x} \big(1 - s_c(x, K_p)\big)^{1 - z_x}. \tag{12}$$

This generalized Bayes update uses partial agreement without thresholding and reduces to binary Bayes at $z_x \in \{0, 1\}$ (Bissiri et al., 2016). Agreement alone cannot establish correctness: identical predictions may both be wrong. We recompute the coverage–retention score with $p_{m,t}(c \mid z_x)$ to select $K^\star(x, m, t)$:

$$K^*(x, m, t) = \arg \max_{K \in \{K_0(x), \ldots, N_t\}} \sum_{r=1}^{K} q_\phi(r \mid x) \sum_{c \geq K - r} p_{m,t}(c \mid z_x). \tag{13}$$

The reader then makes one final prediction using the selected pool.

5 Experimental Design

Protocol and readers. We evaluate ATIS, SNIPS, Banking77, and CLINC150 (Hemphill et al., 1990; Coucke et al., 2018; Casanueva et al., 2020; Larson et al., 2019) under Blend/Mix (Qin et al., 2020; Yoon et al., 2024), yielding eight leave-one-domain-out settings. Training and model selection use only the other three domains. The complete target test split supplies unlabeled coverage posteriors for calibration-call allocation; target labels are used only after actions and predictions are fixed.

Nine frozen local readers cover Qwen3 0.6B/1.7B/4B/8B/14B/32B and Llama 1B/3B/8B (Qwen Team, 2025b; Dubey et al., 2024). The source response model is fitted using outputs from Qwen3 0.6B/4B/14B. The remaining six readers evaluate unseen-size and cross-family transfer. Qwen3 14B/32B use 4-bit loading; other local readers use bfloat16 (BF16). Table 4 reports two output-only application programming interfaces (APIs).

Baselines and supervision. Exposure comparisons share the reader, prompt, descriptions, and cosine ranking: Full, fixed $K \in \{5, 10, 20\}$, reader-blind Adaptive-k, and reader-specific LDAR. Fixed depths are capped by schema size, with Full representing duplicates; Appendix D.6 reports 99.53% agreement for separate-run duplicates. LDAR uses reader-specific outcomes from 64 training and 32 development utterances per source domain. CAPE trains coverage on complete source splits

Table 1: **Local F1/Exact (%).** All local averages 72 settings; fixed depths are capped by schema size. [a]Top-5: Parikh et al. (2023). [b]Recommended $K = 10$ and sweep endpoint $K = 20$: Hong et al. (2024). [c]Adaptive-k: Taguchi et al. (2025). [d]Source-only LDAR: Shim et al. (2026). †Source response-geometry reader; ‡4-bit loading.

Reader / target transfer	Full	$K = 5$[a]	$K = 10$[b]	$K = 20$[b]	Adaptive-k[c]	LDAR[d]	CAPE
Panel A: By reader; average over four target transfers and Blend/Mix							
Seen Qwen readers (source response geometry)							
Qwen3-0.6B†	51.81/18.61	52.49/17.68	50.80/14.74	50.11/18.31	50.39/13.08	53.32/18.58	57.88/28.12
Qwen3-4B†	79.39/61.97	74.53/53.66	79.19/60.52	80.51/62.97	78.18/58.88	76.65/55.21	79.22/60.92
Qwen3-14B†,‡	84.83/65.96	78.34/56.37	82.51/61.91	84.10/64.46	81.99/61.38	83.87/64.19	83.39/64.76
Seen Qwen average	**72.01/48.85**	**68.45/42.57**	**70.83/45.72**	**71.57/48.58**	**70.19/44.45**	**71.28/46.00**	**73.50/51.27**
Unseen Qwen readers (held-out sizes)							
Qwen3-1.7B	58.38/31.66	60.62/32.85	60.90/32.39	60.96/32.29	59.99/32.34	56.53/27.23	63.28/34.67
Qwen3-8B	79.24/61.21	75.55/54.32	79.94/61.29	80.49/62.67	78.78/59.50	78.92/59.45	79.61/60.71
Qwen3-32B‡	84.73/67.10	78.04/55.58	82.54/62.34	84.14/65.51	81.78/61.22	83.66/64.96	83.69/65.95
Unseen Qwen average	**74.12/53.33**	**71.40/47.59**	**74.46/52.01**	**75.19/53.49**	**73.52/51.02**	**73.03/50.55**	**75.53/53.78**
Unseen Llama readers (held-out family)							
Llama-3.2-1B	32.34/2.51	50.70/2.85	45.50/2.15	40.78/2.00	45.74/2.20	53.69/17.85	56.56/21.62
Llama-3.2-3B	58.00/17.76	60.93/25.74	60.86/21.88	60.81/24.04	61.21/24.81	58.80/21.63	64.85/32.51
Llama-3.1-8B	65.32/25.82	65.38/29.34	68.79/30.83	67.02/27.79	67.44/30.67	61.74/21.77	67.97/34.94
Unseen Llama average	**51.89/15.36**	**59.00/19.31**	**58.38/18.29**	**56.20/17.94**	**58.13/19.23**	**58.08/20.42**	**63.13/29.69**
Panel B: By target/schema transfer; average over nine local readers and Blend/Mix							
SNIPS (7)	79.94/54.80	80.64/55.22	79.93/54.77	79.93/54.77	79.86/54.54	75.73/47.92	84.02/62.80
ATIS (18)	66.38/35.56	62.89/29.87	66.95/34.13	66.29/35.50	66.23/34.47	66.92/36.44	67.26/37.46
Banking77 (77)	52.18/26.45	54.82/25.01	56.11/26.20	54.95/27.08	55.21/25.73	57.16/28.54	59.18/31.66
CLINC150 (147)	65.51/39.90	66.80/35.86	68.58/39.58	69.46/42.67	67.81/38.19	70.05/43.04	72.42/47.73
All local	**66.00/39.18**	**66.29/36.49**	**67.89/38.67**	**67.66/40.01**	**67.28/38.23**	**67.46/38.99**	**70.72/44.91**

without source outcomes from held-out readers, so this is not an equal-data efficiency comparison. All methods cover 72 local and 16 hosted settings. Appendix L evaluates Direct/DSCP × Full/CAPE across 72 local settings, refitting the source response model for each procedure.

For local matched-budget comparisons, reader-calibrated coverage thresholding selects $K_\tau(x) = \min\{K : C_\phi(K \mid x) \geq \tau\}$ from the same coverage posterior. Each reader–schema pair chooses τ by synthetic Exact; inference uses three-view aggregation.

Metrics and controls. We report sample-F1 and Exact as macro-averages over target–suite settings, together with exposure, gold coverage, reader calls, and tokens. Gold coverage is measured after inference. CAPE uses two order views and one final prediction per query, in addition to schema calibration. Three-view controls cover all 72 local settings, and Appendix Table 15 reports three training seeds.

The pre-query audit uses 400 prespecified queries per target–suite pair (3,200 total), primarily on the six held-out readers. References share evaluation queries and reader outputs; the reader-blind reference uses cross-fitted target labels to estimate probabilities. Appendix M specifies metrics, depth grids, cross-fitting, and bootstrapping. Headline results use complete test splits.

6 RESULTS

Local-Reader Performance and Transfer Table 1 shows the highest local macro F1 and Exact for CAPE, although it trails Full on several larger Qwen readers (Table 11). Increasing LDAR's training set from 192 to 768 utterances leaves a 5.03-point mean Exact advantage for unchanged CAPE across two readers (Table 25), concentrated in Llama-3.2-3B; the Qwen3-8B difference remains uncertain.

Matched Calibrated Policies Table 2 shows that CAPE outperforms reader-calibrated coverage thresholding in both the all-reader and held-out-reader aggregates under matched budgets. CAPE basic omits Fisher/EVI selection and query updating and uses three-view aggregation. It improves

Table 2: **Matched calibrated policies.** Budgets are up to 64 synthetic classification calls per reader–schema pair and three calls per query. Exact averages 72 settings (All nine) or 48 (Held-out six); mean K averages all 72. Paired 95% query-bootstrap CIs hold models, readers, and calibration fixed.

Policy / comparison	All nine	Held-out six	Mean K
Exact (%)			
K_τ + 3-view	44.05	40.83	12.80
CAPE basic	44.42	41.21	10.85
CAPE	**44.91**	**41.73**	11.27
Paired ΔExact [95% CI] (pp)			
CAPE − (K_τ + 3-view)	+0.86 [+0.61, +1.15]	+0.90 [+0.32, +1.48]	—
CAPE − CAPE basic	+0.49 [+0.22, +0.78]	+0.52 [+0.14, +0.95]	—

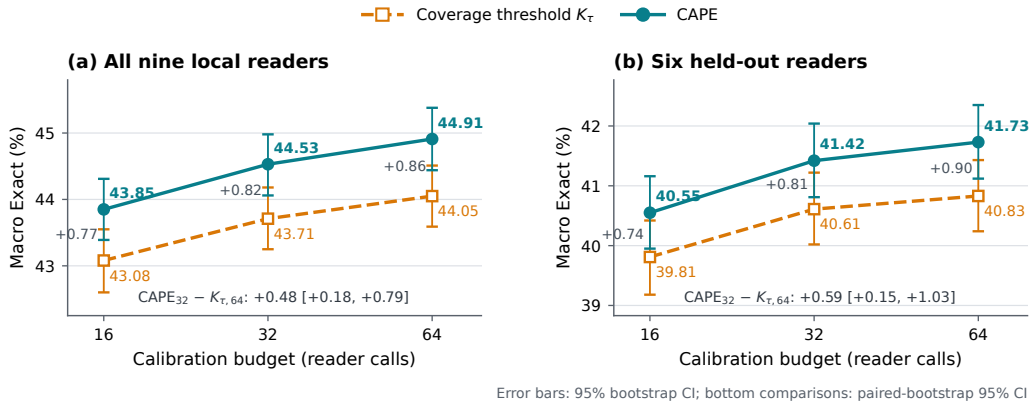


Figure 3: **Calibration-call efficiency.** Macro Exact over 72 all-reader (a) and 48 held-out-reader (b) settings. Budgets count synthetic classification calls per reader–schema pair, excluding naturalization; both policies use three query calls. Bars and bottom comparisons show 95% bootstrap CIs; bottom comparisons are paired.

over thresholding by 0.37 Exact points [0.09, 0.66] overall; the full procedure yields further gains in both scopes.

Figure 3 further shows that 32-call CAPE exceeds 64-call thresholding in both aggregates, counting synthetic classification calls.

Table 3: **Adaptation ablations.** Means cover 72 settings, except $\text{Exact}_{\text{unseen}}$ (48 held-out-reader settings) and large-schema Exact (36 Banking77/CLINC150 settings). Cov. is complete gold-set inclusion, measured post hoc. Shuffling changes reader-posterior assignments within each target–suite, fixing queries, rankings, and tested readers.

Variant	Reader adapt.	Query adapt.	F1	Exact	$\text{Exact}_{\text{unseen}}$	Large-schema Exact	$\bar{K}$	Cov.
Stage 1 only	–	–	66.20	36.98	35.81	34.22	3.77	49.91
+ Reader adaptation	✓	–	69.65	43.35	40.82	38.88	8.71	66.58
+ Query adaptation	–	✓	69.68	42.46	38.93	38.34	9.22	71.78
Full CAPE	✓	✓	**70.72**	**44.91**	**41.73**	**39.70**	11.27	71.99
Shuffled reader posterior	✓	✓	69.95	43.40	40.53	38.86	10.37	72.97

Reader- and Query-Level Adaptation Table 3 shows that reader adaptation and query adaptation raise Exact from 36.98 with Stage 1 alone to 43.35 and 42.46, respectively. Combining both reaches 44.91. Reader calibration provides a reader-specific capacity distribution, which query-level agreement then refines for each input. Shuffling reader posteriors lowers Exact despite higher coverage, further supporting reader-specific calibration. All nine readers show positive Exact differences over both matched controls on Banking77 and CLINC150 in Table 27. Larger-pool cases contribute 1.07 percentage points to the net gain over thresholding in Table 18.

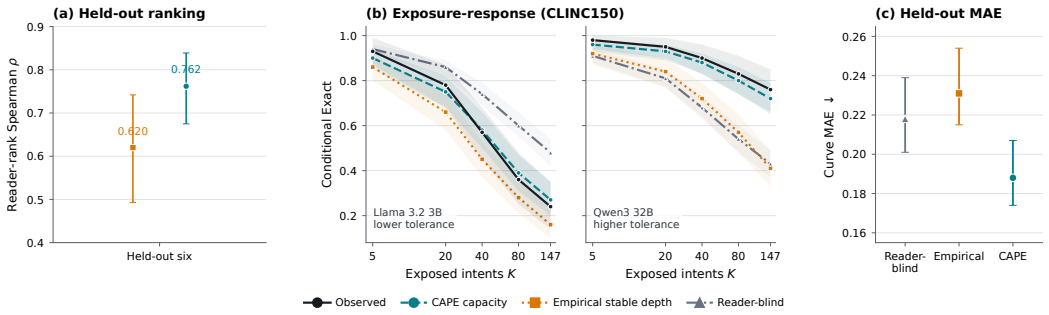


Figure 4: **Pre-query response prediction on six held-out readers.** (a) Reader-rank correlation macro-averaged over target–suite–depth conditions. (b) CLINC150 retention for readers selected by pre-query capacity, on shared anchor-exact queries with $K_0 < 5$ (64 Blend/72 Mix). (c) Macro curve MAE. Intervals are 95% query-bootstrap CIs conditional on fixed readers and calibration. Reader-blind uses cross-fitted target labels; see Appendix M.

Predicting Unseen Reader Responses Figure 4 evaluates pre-query retention curves on six held-out readers, measuring correctness above K_0 conditional on an exact anchor response. CAPE raises reader-rank correlation from 0.620 to 0.762 over the empirical reference and lowers curve mean absolute error (MAE) relative to both references (Tables 22–23). Individual-query Brier scores (Brier, 1950) and negative log-likelihood (NLL) improve over the empirical reference, with both 95% CIs below the constant-0.5 scores. The target-supervised reader-blind reference achieves lower losses. Appendix M details the matched comparisons and cross-fitting.

We next test whether capacity estimates improve exposure decisions, using all sampled queries. Against the reader-blind reference, the capacity-based policy reduces regret relative to the best observed query-grid action from 3.10 to 1.95 points and raises Exact from 40.05 to 41.20. Only this pre-query audit fixes anchor retention to one; deployment uses the full coverage–retention surrogate.

To check the monotone-boundary approximation, Table 24 measures recovery after an earlier error. Recovery occurs in 47 of 860 eligible held-out trajectories, or 5.47% [4.38, 6.71], confirming that observed responses can be non-monotone.

Table 4: **Hosted-reader F1/Exact (%).** Per-API means cover eight all-target or four large-schema (Banking77/CLINC150) settings. $\bar{K}_{\mathrm{CAPE}}$ is mean exposure; bold marks the best in each scope. Table 12 gives target-level results.

Reader	Scope	Full	Adaptive-k	LDAR	CAPE	$\bar{K}_{\mathrm{cape}}$
GPT-4o-mini	All targets	85.89/70.84	82.26/63.18	**85.99/70.86**	84.91/68.91	21.08
	Large schemas ($N \geq 77$)	83.14/66.80	76.64/53.26	83.35/67.12	**83.70/67.31**	36.49
Gemini-3.1-Flash-Lite	All targets	**89.76/77.90**	85.05/67.29	89.30/76.68	88.58/75.63	21.66
	Large schemas ($N \geq 77$)	88.16/76.40	79.84/57.09	88.16/76.27	**88.65/76.92**	37.57

Hosted Readers and Reasoning Procedures Table 4 shows that CAPE outperforms Adaptive-k on both hosted readers but trails Full and LDAR overall. Small gains over Full and LDAR are limited to the Banking77/CLINC150 aggregate, with mean exposure reduced from 112 to 37.03 intents relative to Full.

Table 21 tests whether exposure gains persist with query decomposition. Across 72 local settings, CAPE improves Exact over Full by 5.73 points under Direct and 5.72 under DSCP. DSCP+CAPE reaches 47.85 Exact, supporting the complementary roles of adaptive exposure and the reasoning procedure.

Independent naturalization and raw descriptions preserve capacity rankings (mean $\rho = 0.946$–0.975) and macro Exact (44.76–44.91; Appendix I.6).

Table 5: **Inference cost and accuracy.** F1, Exact, and $\bar{K}$ average 72 local settings. Full uses one call; the other controls aggregate three views. CAPE uses two evidence calls and one final prediction. Tokens/q is query-weighted and includes amortized schema costs.

Method	Calls/q	$\bar{K}$	Tokens/q	F1	Exact
Full	1	62.25	1,869	66.00	39.18
$K = 20$	3	16.25	1,739	67.49	42.18
Adaptive-k	3	11.82	1,450	67.31	40.01
LDAR	3	26.96	2,772	67.71	40.26
CAPE	3	11.27	2,017	**70.72**	**44.91**

Inference Cost Table 5 compares inference cost and accuracy. CAPE exceeds three-view $K = 20$ aggregation by 2.73 Exact points. Including amortized schema costs, CAPE uses 2,017 tokens per query, versus 1,869 for Full and 1,739 for three-view $K = 20$.

Schema Exposure and Error Analysis Tables 1 and 13 show higher macro Exact for CAPE than Adaptive-k and LDAR despite lower gold coverage.

Let G denote Full errors corrected by CAPE, L_{cov} Full-correct predictions lost through gold omission, and L_{covered} Full-correct predictions lost despite gold coverage. Their disjoint rates satisfy

$$\Delta\text{Exact} = G - L_{\text{cov}} - L_{\text{covered}}. \tag{14}$$

Table 6: **Full-to-CAPE Exact transitions.** Query rates are macro-averaged over 72 settings; 95% CIs use 10,000 setting-bootstrap replicates. Labels are used only for this post-hoc audit.

Transition	Rate (pp)	95% CI
Gain: Full wrong → CAPE exact	12.02	[10.26, 13.90]
Coverage loss: Full exact → CAPE wrong, gold missing	3.85	[3.05, 4.88]
Covered-pool loss: Full exact → CAPE wrong, gold covered	2.44	[2.15, 2.75]
Net Δ Exact	+5.74	**[**+3.52**,** +7.88**]**

Gold omission dominates Full-to-CAPE losses (Tables 6 and 26). On ATIS, predicted coverage averages 88.9% versus 67.2% observed (Table 28). At fixed depths and predictions, gold-rank substitution reduces ATIS utility Brier from 0.365 to 0.168; this diagnostic uses target labels (Appendix O). Full-default arbitration retains Full unless a verifier confirms the pre-query proposal. This supplementary variant raises ATIS Exact from 37.46 to 39.51, versus 38.71 for three-view Full. Both recover the deficits against Full for Qwen3-4B/8B/14B/32B; their relative gains vary by reader (Appendix P).

7 Discussion and Limitations

Gold omission and small-schema hosted performance remain limitations. Calibration can fail when synthetic and real queries differ or incorrect predictions agree. Online use requires a representative unlabeled buffer or source-reference allocation. Self-tests require faithful descriptions and reader calls. The source response model uses three Qwen readers, and evaluation queries contain at most three intents. Naturalizer and DSCP controls cover only the tested interventions. Four-bit loading confounds size and precision for the two largest Qwen readers; Appendix J reports a matched 8B control. Unequal source budgets and reader supervision limit conclusions about equal-data efficiency and LDAR saturation. Supplementary arbitration recovers the larger Qwen readers' ATIS losses but lowers overall Exact by 0.93 points relative to native CAPE. The coverage–retention framework may also inform tasks that select from described candidates, such as entity linking and tool selection; evaluating this transfer remains future work.

8 Conclusion

CAPE combines coverage prediction, reader calibration, and query-level agreement to select exposure without target annotations or reader internals. Calibration predicts conditional retention on held-out readers, and CAPE improves Exact over reader-calibrated coverage thresholding under matched budgets.

AI USE STATEMENT

We used AI tools to assist with writing and improve the clarity and readability of the manuscript. The use of language models to generate synthetic calibration utterances is described in the methodology and appendices. The authors take full responsibility for the final content of this paper.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

Jan De Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32:1–24, 2010.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Jinjie Fang, Runwen You, Wanli Shi, Wenkang Wang, Ganyu Wang, Haozhen Zhang, Yi Chang, and Bin Gu. Online black-box prompt optimization with regret guarantees under noisy feedback. In *International Conference on Learning Representations*, 2026.

Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

Taesuk Hong, Youbin Ahn, Dongkyu Lee, Joongbo Shin, Seungpil Won, Janghoon Han, Stanley Jungkyu Choi, and Jungyun Seo. Exploring the use of natural language descriptions of intents for large language models in zero-shot intent classification. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 458–465, 2024. doi: 10.18653/v1/2024.sigdial-1.39.

Ruoyu Hu, Foaad Khosmood, and Abbas Edalat. Exploring description-augmented dataless intent classification. In *Proceedings of the 6th Workshop on NLP for Conversational AI*, pp. 13–36, 2024. URL https://aclanthology.org/2024.nlp4convai-1.2/.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, pp. 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. Exploring zero and few-shot techniques for intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Industry Track*, pp. 744–751, 2023. doi: 10.18653/v1/2023.acl-industry.71.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1807–1816, 2020. doi: 10.18653/v1/2020.findings-emnlp.163.

Libo Qin, Qiguang Chen, Jingxuan Zhou, Jin Wang, Hao Fei, Wanxiang Che, and Min Li. Divide-solve-combine: An interpretable and accurate prompting framework for zero-shot multi-intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25038–25046, 2025. doi: 10.1609/aaai.v39i23.34688.

Qwen Team. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025b.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, 1960.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Seongwoong Shim, Myunsoo Kim, Jae Cho, and Byung-Jun Lee. Beyond rag vs. long-context: Learning distraction-aware retrieval for efficient knowledge grounding. In *International Conference on Learning Representations*, volume 2026, pp. 122532–122565, 2026.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7967–7977, 2022.

Mujeen Sung, James Gung, Elman Mansimov, Nikolaos Pappas, Raphael Shu, Salvatore Romeo, Yi Zhang, and Vittorio Castelli. Pre-training intent-aware encoders for zero-and few-shot intent classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10433–10442, 2023.

Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. Efficient context selection for long-context QA: No tuning, no iteration, just adaptive-k. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20105–20130, 2025. doi: 10.18653/v1/2025.emnlp-main.1017.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pp. 5–42, 2011.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G. M. Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *International Conference on Learning Representations*, 2025.

Yejin Yoon, Jungyeon Lee, Kangsan Kim, Chanhee Park, and Taeuk Kim. BlendX: Complex multi-intent detection with blended patterns. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2428–2439, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.218/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Supplementary Material

A Derivation and Decision Rule . . . 14
A.1 Exact-Set Probability . . . 14
A.2 Why No Fixed Depth Is Generally Optimal . . . 14
A.3 The Coverage–Distraction Balance . . . 14
A.4 Posterior Decision Rule . . . 15
A.5 Shape Constraint for Capacity Survival . . . 15
B Architecture and Source Training . . . 15
B.1 Schema and Query Representations . . . 15
B.2 Incremental Item Difficulty . . . 16
B.3 Source Training Objectives . . . 17
B.4 Source Response-Geometry Reader Panel . . . 18
C Test-Time Adaptation Details . . . 18
C.1 Schema-Derived, Label-Free Self-Test Construction . . . 18
C.2 Reader Posterior . . . 18
C.3 Query Probe Selection . . . 20
C.4 Deterministic Same-Pool Perturbation . . . 20
C.5 Generalized Bayes Update . . . 21
C.6 End-to-End Algorithm . . . 21
D Data, Prompts, Readers, and Baselines . . . 21
D.1 Dataset Statistics . . . 21
D.2 Deployment-Query Availability . . . 21
D.3 Frozen-Reader Prompt . . . 22
D.4 Naturalization Prompt . . . 22
D.5 Readers . . . 22
D.6 Baseline Definitions . . . 22
E Full Performance Results . . . 24
E.1 Full Local Reader–Target Matrix . . . 24
E.2 Target-Wise Hosted-API Results . . . 25
F Exposure Behavior . . . 25
G Source-Reader Robustness . . . 25
H Training-Seed Stability . . . 25
I Mechanism and Sensitivity Controls . . . 26
I.1 Main-Component Ablation Definitions . . . 27
I.2 Simple Calibrated Policy Controls . . . 29
I.3 Empirical Reader-Calibration Control . . . 31
I.4 Fixed versus EVI-Selected Query Probes . . . 32
I.5 Reader-Posterior Swap . . . 33
I.6 Naturalization Artifact Controls . . . 33
I.7 Shape and Numerical Checks . . . 33
J Matched-Precision Control . . . 33
K Cost and Amortization . . . 34
L Reasoning-Procedure Robustness . . . 34
L.1 Strict DSCP $\times$ Exposure Protocol . . . 35
M Predictive Validity of Pre-Query Reader Calibration . . . 35
M.1 Evaluation Grid and Retention Event . . . 36
M.2 Diagnostic References . . . 36
M.3 Controlled Selection and Regret . . . 37
M.4 Example-Curve Cohort and Uncertainty . . . 37
M.5 Observed Recovery on Real-Query Trajectories . . . 38
M.6 Setting-Level Exact-Transition Decomposition . . . 39
N LDAR Source-Training Budget Sensitivity . . . 39
O Reader-Wise Gains and ATIS Selection Diagnostics . . . 41
O.1 Gains by Reader and Target . . . 41
O.2 Coverage and Retention at the Selected Depth . . . 41
O.3 Gold-Rank Substitution at Fixed Policy Outputs . . . 42
P Recovering ATIS Losses with Full-Default Arbitration . . . 43

A Derivation and Decision Rule

A.1 Exact-Set Probability

For a target query x, let Y be its unknown gold intent set and π_x the cosine-similarity ranking of the target schema. Define the minimum cover rank R as in Eq. 2. The event that a prefix contains the complete gold set is $A_K = [R \le K]$. The event that the reader predicts exactly is $E_K = [\widehat{Y}_K = Y]$.

Lemma 1 (coverage–preservation factorization). For every query, reader, and prefix depth,

$$\Pr(E_K = 1 \mid x, m) = \Pr(A_K = 1 \mid x) \Pr(E_K = 1 \mid A_K = 1, x, m). \tag{15}$$

Proof. If $A_K = 0$, at least one gold intent is absent from the allowed output schema. The constrained reader cannot emit this intent, so $\Pr(E_K = 1, A_K = 0 \mid x, m) = 0$. Therefore

$$\Pr(E_K = 1 \mid x, m) = \Pr(E_K = 1, A_K = 1 \mid x, m) \tag{16}$$

$$= \Pr(A_K = 1 \mid x) \Pr(E_K = 1 \mid A_K = 1, x, m). \tag{17}$$

No conditional-independence assumption is used. □

The first term measures retrieval coverage. The second reflects the reader's reasoning, sensitivity to distractors, output formatting, and capacity. We approximate the second term using synthetic probes constructed from the schema. These probes measure correctness after complete gold coverage as the number of additional intents increases, without requiring target annotations. This schema-level estimate serves as a transferable proxy. It does not identify the true conditional preservation probability for each target utterance, so the decision rule remains a utility surrogate.

A.2 Why No Fixed Depth Is Generally Optimal

Let $U_{x,m}(K) = \Pr(E_K = 1 \mid x, m)$ and let $\mathcal{K} = \{1, \dots, N\}$.

Proposition 1 (fixed-depth regret). If there exist two reader–query pairs (x_1, m_1) and (x_2, m_2) with unique optimal depths $K_1^* \neq K_2^*$, then every global fixed depth K has positive regret on at least one pair:

$$\max_{j \in \{1,2\}} \left[U_{x_j,m_j}(K_j^*) - U_{x_j,m_j}(K) \right] > 0. \tag{18}$$

Proof. A single K cannot equal both distinct unique maximizers. On the pair for which $K \neq K_j^*$, uniqueness implies strict positive regret. □

This proposition is elementary by design. Its empirical premise is the substantive question. Figure 1 provides finite-grid evidence for that premise; it is a deployment-oriented diagnostic rather than an exhaustive oracle claim.

A.3 The Coverage–Distraction Balance

Write $C_K = C_\phi(K \mid x)$ and let $S_{m,t}(d)$ be the estimated probability that the reader tolerates at least d excess intents after gold coverage. The implemented surrogate is

$$\widetilde{U}_K = \sum_{r=1}^{K} q_\phi(r \mid x) S_{m,t}(K - r). \tag{19}$$

For $C_K > 0$, define the induced conditional retention proxy $\bar{S}_K = \widetilde{U}_K / C_K$, so that $\widetilde{U}_K = C_K \bar{S}_K$. At adjacent depths for which C_K, C_{K+1}, and $\bar{S}_{K+1}$ are positive, expanding from K to $K+1$ increases this surrogate exactly when

$$\frac{C_{K+1}}{C_K} > \frac{\bar{S}_K}{\bar{S}_{K+1}}. \tag{20}$$

The left side is the multiplicative coverage gain. When $\bar{S}_{K+1} < \bar{S}_K$, the right side is the corresponding retention cost; the underlying state survival $S_{m,t}(d)$ is non-increasing even though the mixture $\bar{S}_K$ can also change when newly covered rank mass enters at K. Equation 20 explains three local surrogate behaviors without choosing a global reference depth:

- **Expansion is harmful** when coverage has nearly saturated but induced retention proxy decreases, so the inequality fails at the current depth.
- **Expansion is useful** when substantial cover-rank mass remains and the reader tolerates added descriptions, so the inequality holds.
- **Expansion leaves the surrogate unchanged** when the multiplicative coverage gain equals the multiplicative retention cost.

A.4 Posterior Decision Rule

Let $q_\phi(r \mid x)$ be the required-rank posterior and let $p_{m,t}(c)$ be the discrete effective-capacity posterior induced for reader m under target schema t. The implementation indexes capacity states by $c \in \{-1, 0, \ldots, N_t - 1\}$, with -1 denoting failure even when the gold set is first covered. Its survival tail is $S_{m,t}(d) = \sum_{c \geq d} p_{m,t}(c)$. The posterior-informed coverage–retention surrogate is

$$\widetilde{\mathcal{U}}(K \mid x, m, t) = \sum_{r=1}^{K} q_\phi(r \mid x) S_{m,t}(K - r) = \sum_{r=1}^{K} q_\phi(r \mid x) \sum_{c \geq K-r} p_{m,t}(c). \tag{21}$$

The capacity tail is estimated from synthetic rather than target utterances. We therefore use Eq. 21 as a decision objective without equating it to $\Pr(E_K = 1 \mid x, m)$. The implementation evaluates it for every $K \in \{K_0, \ldots, N\}$ and returns the exact discrete argmax. It does not stop at a score threshold and imposes no fixed upper pool size.

Its diagnostic mean is $\hat{c}_{m,t} = \sum_c c\, p_{m,t}(c)$, as in Eq. 8. The code evaluates Eq. 21 by fast Fourier transform (FFT) convolution; this is algebraically equivalent to the direct sum and only changes runtime.

A.5 Shape Constraint for Capacity Survival

Target calibration represents each synthetic probe by a latent largest-exact boundary, which is an explicit monotone-capacity approximation. Aggregating a finite set of interval posteriors over different risk sets can nevertheless produce a non-monotone raw survival estimate $\widetilde{S}_d$. For excess depth d, let w_d be the number of probes for which that depth is available within the schema. Let $\widetilde{S}_d$ be the raw survival estimate, and let $S = (S_0, \ldots, S_{N-1})$ be the candidate survival curve. We obtain the fitted curve S^* by weighted isotonic regression:

$$S^* = \operatorname*{arg\,min}_{1 \geq S_0 \geq \cdots \geq S_{N-1} \geq 0} \sum_{d=0}^{N-1} w_d (S_d - \widetilde{S}_d)^2. \tag{22}$$

We solve Eq. 22 using the pool-adjacent-violators algorithm (PAVA) (De Leeuw et al., 2010). This projection makes the inferred capacity distribution valid and leaves an already monotone estimate unchanged. It is a modeling constraint on the schema-level tolerance summary, not a claim that every raw reader trajectory is monotone; Section 6 evaluates the monotone prediction against independently collected endpoint outcomes. Appendix I reports its operational consequence by comparing the complete decision pipeline with raw and projected capacity survival.

B Architecture and Source Training

B.1 Schema and Query Representations

All experiments use the frozen 768-dimensional SentenceTransformer checkpoint `sentence-transformers/all-mpnet-base-v2`, resolved to Hugging Face revision `e8c3b32edf5434bc2275fc9bab85f82640a19130`. The encoder remains in evaluation mode. We take the first three descriptions for each intent. Underscores and hyphens in an intent name are replaced by spaces before encoding. The encoder produces unit-normalized vectors for the utterance $u = g(x)$, normalized intent name n_i, each individual description, and the combined schema text e_i, whose literal template is `Intent: {name}. Description: {description 1} {description 2} {description 3}`. The description aggregate

d_i is the arithmetic mean of the three individually normalized description vectors. Retrieval uses $s_i = u^\top e_i$. In the following equations, i indexes the intent's position after cosine ranking. Let $\widetilde{u}$ be a learned projection of the query embedding u. We compute a schema representation a_i, a rank-feature representation v_i, and their fused input $h_i^{(0)}$ using multilayer perceptrons (MLPs). Here $[\,;\,]$ denotes concatenation, $|\cdot|$ elementwise absolute value, and $\odot$ elementwise multiplication:

$$a_i = \mathrm{MLP}_{\mathrm{schema}}[n_i; d_i; |n_i - d_i|; n_i \odot d_i], \tag{23}$$

$$v_i = \mathrm{MLP}_{\mathrm{rank}}[i/N; \log(1+i); s_i; s_{i-1} - s_i; s_i - s_{i+1}], \tag{24}$$

$$h_i^{(0)} = \mathrm{MLP}_{\mathrm{pair}}[\widetilde{u}; a_i; |\widetilde{u} - a_i|; \widetilde{u} \odot a_i; v_i], \tag{25}$$

A two-layer Transformer encoder with four heads maps the fused inputs to contextual representations h_i. Let ℓ_i be the logit for choosing rank i as the minimum gold-cover rank:

$$\ell_i = \mathrm{MLP}_{\mathrm{cut}}[h_i; \widetilde{u}; i/N; \log(1+i)], \tag{26}$$

Collecting the logits into $\ell = (\ell_1, \ldots, \ell_N)$ gives $q_\phi(i \mid x) = \mathrm{softmax}(\ell)_i$.

Table 7: Coverage-selector architecture. D is the sentence-embedding dimension and $H = 256$. Abbreviations are layer normalization (LN), Gaussian error linear unit (GELU), and feed-forward network (FFN).

Component	Input dimension	Structure
Schema fusion	$4D$	Linear–LN–GELU–Dropout–Linear
Utterance projection	D	Linear–LN–GELU
Rank projection	5	Linear–LN–GELU
Pair projection	$5H$	Linear–LN–GELU–Dropout–Linear
Set encoder	H	2 Transformer layers, 4 heads, FFN $4H$
Cutoff head	$2H + 2$	Linear–LN–GELU–Linear–GELU–Linear

The relative-rank feature avoids a learned maximum-position table and permits evaluation on target schemas larger than those seen in source training.

B.2 Incremental Item Difficulty

Let $\bar{h}_K = K^{-1} \sum_{j \le K} h_j$ be the prefix representation and $\bar{h}_{K_0}$ the anchor representation. Each rank receives

$$r_K = \mathrm{MLP}_{\mathrm{item}}[\bar{h}_K; \bar{h}_{K_0}; \bar{h}_K - \bar{h}_{K_0}; h_K]. \tag{27}$$

The final model maps r_K to a per-addition logistic difficulty d_K. For a reader tolerance θ, the modeled conditional probability that the upper endpoint is exact, given exactness at the lower endpoint, is

$$G_\theta(U \mid L, x) \coloneqq \Pr_\psi(E_U = 1 \mid E_L = 1, x, m) = \prod_{j=L+1}^{U} \sigma(\theta - d_j(x)). \tag{28}$$

This follows item response theory (IRT): reader tolerance and added-item difficulty interact on one latent coordinate (Rasch, 1960). Equivalently,

$$\log G_\theta(U \mid L, x) = \sum_{j=L+1}^{U} \log \sigma(\theta - d_j(x)), \tag{29}$$

so each added rank contributes a non-positive increment to modeled log retention. This parameterization describes the effect of cumulative exposure, rather than applying the chain rule to observed exactness events. Specifically, Eq. 28 models the upper-endpoint event E_U, without requiring every event in $\bigcap_{j=L+1}^{U} \{E_j = 1\}$ to hold. An observed prediction may be incorrect at an intermediate depth and exact at a larger depth. Monotonicity therefore constrains endpoint probabilities without assuming permanent failure after an error.

At deployment, the item-response model is used only to design informative schema probes. The new reader's posterior is estimated directly from the resulting target-schema outcomes (Appendix C.2), without fitting a new θ for Eq. 28.

B.3 Source Training Objectives

B.3.1 Coverage Training

For each source utterance, we compute the exact minimum gold-cover rank from the immutable cosine ranking and minimize Eq. 4. Domains are balanced within an epoch. The selected checkpoint minimizes the macro source-validation cross-entropy, so a large source dataset cannot dominate selection. The released runs apply no source-example cap: they use the full training and development splits reported in Table 9.

B.3.2 Reader and Expansion Geometry

For a source reader m and source synthetic query (x, Y), let $\mathcal{D}_x$ and $L_x = \min \mathcal{D}_x$ be the prespecified observed depths from Appendix B.4, all of which are at least the minimum gold-cover rank $R(x, Y)$. Let $a_{xm} = E_{L_x} \in \{0, 1\}$ indicate exact prediction at the smallest observed source endpoint. An anchor discrimination $\alpha_{x,L_x} > 0$ and difficulty β_{x,L_x} give

$$\Pr(a_{xm} = 1) = \sigma(\alpha_{x,L_x}\theta_m - \beta_{x,L_x}). \tag{30}$$

For every observed pair $L < U$ whose lower endpoint is exact, define $b_{xmLU} = [E_U = 1]$. Its modeled probability is Eq. 28. Intermediate outcomes are not conjoined: an exact–wrong–exact trajectory gives $b_{xmLU} = 1$ for the endpoint pair (L, U). Let $\mathcal{A}$ contain the source query–reader anchor observations (x, m), and let $\mathcal{B}$ contain the observed endpoint pairs (x, m, L, U) whose lower endpoint is exact. Let p_ψ denote the Bernoulli probabilities assigned by the anchor and expansion models to their observed binary outcomes. The response model minimizes the negative conditional composite log-likelihood:

$$\mathcal{L}_{\text{IRT}} = -\frac{1}{|\mathcal{A}| + |\mathcal{B}|}\left[\sum_{(x,m)\in\mathcal{A}} \log p_\psi(a_{xm}) + \sum_{(x,m,L,U)\in\mathcal{B}} \log p_\psi(b_{xmLU} \mid E_L = 1, x, m)\right]. \tag{31}$$

Pairs derived from the same query reuse reader outputs and need not be independent; Eq. 31 is therefore a composite likelihood (Varin et al., 2011), not a joint-independence claim. There is no manually weighted sum: all observations contribute their log-likelihood and normalization is by the total event count. Source reader tolerances are centered and normalized by their root mean square (RMS) for identifiability. The coverage model is frozen while the item-response residual head is fitted.

Table 8: Optimization settings used for every leave-one-domain-out run. The item-response stage reuses and freezes the trained coverage selector, including its Transformer with two layers and four attention heads. The response head and source-reader tolerance parameters are trained in this stage. Precision uses BF16 automatic mixed precision (AMP).

Hyperparameter	Coverage selector	Item-response stage
Hidden width	256	256
Transformer layers / heads	2 / 4	2 / 4 (frozen)
Dropout	0.1	0.1
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	3×10^{-4}
Weight decay	10^{-4}	10^{-4}
Batch size	128	64
Maximum epochs	30	30
Early stopping patience	5	5
Gradient clipping	1.0	1.0
Precision	BF16 AMP	BF16 AMP
Seed	3407	3407
Checkpoint criterion	macro dev NLL	dev joint NLL
Tie breaker	–	pair Brier score

B.4 Source Response-Geometry Reader Panel

The source response geometry is fit with Qwen3-0.6B, Qwen3-4B, and Qwen3-14B. These readers are spaced across the available Qwen capacity range rather than chosen after target evaluation. Their learned scalar coordinates are nuisance parameters used to identify item difficulty; a new reader never receives one of these IDs. At deployment, only target-schema response outcomes infer its posterior.

Source responses are collected from schema-derived tests rather than sampled natural utterances. For a source schema with N_s intents, we create one test for every anchor intent and each gold cardinality in $\{1, 2, 3\}$, giving $3N_s$ known-answer tests. Each test is evaluated at three distinct prefix depths drawn without replacement from its integer interval $[R(x, Y), N_s]$. The draw is deterministic and approximately uniform: candidates are generated from `SHA1(seed | schema-form | domain-id | attempt)` until three distinct depths are obtained, then sorted. This collection procedure instantiates the set $\mathcal{D}_x$ used in Eq. 30. Let $\mathcal{D}_x = \{K_{x,1} < K_{x,2} < K_{x,3}\}$ denote these observed depths and $L_x = K_{x,1}$. The exact gold-cover rank $R(x, Y)$ defines the lower boundary of the sampling interval but is not forced to appear in $\mathcal{D}_x$. Consequently, the source anchor outcome used below is E_{L_x}, the smallest observed gold-covered endpoint, rather than an unobserved response at $R(x, Y)$. This source quantity is also distinct from the predicted deployment anchor $K_0(x)$. Thus each source schema contributes $9N_s$ endpoint observations per reader in the released runs. Holding out CLINC150, Banking77, ATIS, and SNIPS respectively yields 306/516/693/726 source tests and 918/1,548/2,079/2,178 endpoint observations per reader. Across the three source readers these are 2,754/4,644/6,237/6,534 calls per leave-one-domain-out run; the counts are the same for Blend and Mix, for 40,338 source-reader calls over all eight runs. A seeded SHA-1 partition of the synthetic-test identifier assigns complete test trajectories, including all three depths, to disjoint geometry-training and validation halves.

To test whether any one source reader dominates the learned response geometry, the completed leave-one-source-reader-out controls in Appendix G remove Qwen3-0.6B, 4B, and 14B one at a time, refit only the response geometry, and evaluate each refit over all 72 local settings.

C Test-Time Adaptation Details

C.1 Schema-Derived, Label-Free Self-Test Construction

Each intent has multiple descriptions. For calibration, the final description is held out as a semantic query specification and the remaining descriptions are used in the reader prompt. For each anchor intent, we order other intents by description-embedding similarity. We form known gold sets of cardinalities one through $\min(3, |_T|)$ by taking the anchor and its nearest semantic neighbors. The frozen target reader rewrites each semantic specification into one natural first-person clause; clauses are concatenated for multi-intent probes. The gold set is known from construction, not from a target utterance annotation.

The same reader evaluates each synthetic probe under nested schema prefixes. The source response model is used only to specify informative probe identities and initial depths. Each observed endpoint then narrows the probe's boundary interval, rather than fitting a new source-style reader coordinate. Self-tests contain no target utterances and use no target annotations.

Before calibration, we compute $q_\phi(r \mid x)$ for each query in the unlabeled deployment batch $\mathcal{U}_t$, without calling the target reader. These posteriors are used only to average decision value when allocating the remaining calibration calls. For the headline experiments, $\mathcal{U}_t$ contains the complete held-out test split. For the predictive-validity audit, it contains the prespecified 400-query subset for each target–suite pair. The procedure is therefore label-free and transductive, with access to unlabeled queries in addition to the schema.

C.2 Reader Posterior

For synthetic probe b, let R_b be its known minimum gold-cover rank. We represent its latent largest exact prefix by $B_b \in \{R_b - 1, \ldots, N_t\}$, where $B_b = R_b - 1$ means that the reader is not exact even when the complete gold set first becomes available. Before any reader call, B_b lies in $[R_b - 1, N_t]$. An exact observation at queried depth k updates the lower endpoint to k, while an inexact observation

updates the upper endpoint to $k-1$:

$$(\ell_b, u_b) \leftarrow \begin{cases} (k, u_b), & E_k = 1, \\ (\ell_b, k-1), & E_k = 0. \end{cases} \tag{32}$$

Only unresolved $k \in \{\ell_b + 1, \ldots, u_b\}$ are queried. Conditional on the resulting interval data $\mathcal{D}_b$, the implementation uses the maximum-entropy (Jaynes, 1957) boundary posterior

$$p_b(v) \coloneqq \Pr(B_b = v \mid \mathcal{D}_b) = \frac{[\ell_b \le v \le u_b]}{u_b - \ell_b + 1}. \tag{33}$$

Capacity is measured in excess intents after gold coverage. For depth $d \in \{0, \ldots, N_t - 1\}$, define the finite-schema risk set $\mathcal{B}_d = \{b : R_b + d \le N_t\}$. The raw capacity survival is

$$\widetilde{S}_{m,t}(d) = \frac{1}{|\mathcal{B}_d|} \sum_{b \in \mathcal{B}_d} \sum_{v=R_b+d}^{N_t} p_b(v). \tag{34}$$

Depths with an empty risk set receive zero mass. Applying the weighted PAVA projection in Eq. 22, with $w_d = |\mathcal{B}_d|$, gives the non-increasing survival $S_{m,t}(d)$. The discrete capacity posterior used by every subsequent decision is exactly

$$\begin{aligned} p_{m,t}(-1) &= 1 - S_{m,t}(0), \\ p_{m,t}(c) &= S_{m,t}(c) - S_{m,t}(c+1), \quad 0 \le c < N_t - 1, \\ p_{m,t}(N_t - 1) &= S_{m,t}(N_t - 1). \end{aligned} \tag{35}$$

Thus $S_{m,t}(d) = \sum_{c \ge d} p_{m,t}(c)$ and $\hat{c}_{m,t} = \sum_c c\, p_{m,t}(c)$. Equations 33–35 are the complete transformation from target-reader outcomes to $p_{m,t}(c)$; no target θ_m, Gaussian prior, or Laplace map is used.

Initial probe identities and endpoint pairs are selected deterministically by the frozen source probe-design geometry. For candidate synthetic test b, the lower endpoint is its known cover rank $L_b = R_b$. The geometry's hazard head emits a raw value $\eta_{b,j}$ for each added rank and defines

$$h_b(1) = 0, \qquad h_b(j) = \mathrm{softplus}(\eta_{b,j})\ (j > 1), \qquad H_b(k) = \sum_{j=1}^{k} h_b(j). \tag{36}$$

For $j > 1$, the corresponding base logistic difficulty is $d_{b,j}^{\mathrm{base}} = h_b(j) + \log(1 - \exp[-h_b(j)])$, so $\sigma(-d_{b,j}^{\mathrm{base}}) = \exp[-h_b(j)]$ at the centered reference tolerance; the anchor entry is set to zero. Probe design uses the reference-matched auxiliary exponential-hazard model

$$G_\theta^{\mathrm{design}}(U \mid L_b, b) = \exp\!\left\{-e^{-\theta}[H_b(U) - H_b(L_b)]\right\} \tag{37}$$

at $\theta_{\mathrm{ref}} = 0$. At this reference, its retention probability matches that of the product-logistic construction induced by $d_{b,j}^{\mathrm{base}}$. This pointwise probability match does not make the two parameterizations equivalent: their derivatives with respect to θ and hence their Fisher information generally differ. The reference is fixed by centering the source reader coordinates; it is used to rank probes, not assigned to the deployed reader. For every $U > L_b$, write $\Delta H_b(U) = H_b(U) - H_b(L_b)$ and $\widetilde{G}_b(U) = G_0^{\mathrm{design}}(U \mid L_b, b)$. The score

$$\mathcal{J}_b(U) = \left. \frac{[\partial_\theta G_\theta^{\mathrm{design}}(U \mid L_b, b)]^2}{G_\theta^{\mathrm{design}}(U \mid L_b, b)[1 - G_\theta^{\mathrm{design}}(U \mid L_b, b)]} \right|_{\theta=0} = \frac{\widetilde{G}_b(U)\, \Delta H_b(U)^2}{1 - \widetilde{G}_b(U)} \tag{38}$$

is therefore the Bernoulli Fisher information of the auxiliary design model about its scalar coordinate, evaluated at the fixed reference tolerance; it is not the Fisher information of Eq. 28. We first retain, for each test, the upper endpoint maximizing $\mathcal{J}_b(U)$. These candidate pairs are sorted globally by decreasing Fisher score; ties are broken by synthetic-test index and then smaller upper depth. Let $\mathcal{P}_t$ be the resulting set of valid design candidates for target t. A single global rule partitions reader-call budget B as

$$M_t = \min\left(\left\lfloor \frac{B}{2} \right\rfloor, |\mathcal{P}_t| \right), \qquad B_{\mathrm{initial}} = M_t, \qquad B_{\mathrm{adaptive}} = B - M_t. \tag{39}$$

The initial phase evaluates only the lower anchor $L_b = R_b$ once for each of the first M_t probes. The Fisher-optimal upper endpoint is used to score and rank the design candidate; it is not automatically consumed as a second initial call. This design uses the target schema and frozen source geometry but no target-reader output or target label. We allocate the remaining B_{adaptive} calls sequentially. Each unresolved boundary proposes its posterior-median depth. Among these candidates, the next call maximizes the expected increase in mean optimal utility, averaged over the unlabeled queries' coverage posteriors under the current observations. With the default budget $B = 64$, we use 21 initial and 43 adaptive calls for SNIPS, and 32 initial and 32 adaptive calls for ATIS, Banking77, and CLINC150. This split is fixed before target evaluation, without tuning to dataset performance. The budget limits calibration cost but does not cap the final depth K; appendix I reports its sensitivity.

C.3 Query Probe Selection

Before observing query agreement, CAPE chooses the same-pool order probe that maximizes EVI for the final decision. For a candidate probe depth k, the required-rank posterior determines the parameter-free order-stability likelihood

$$s_c(x,k) = \sum_{r=\max(1,k-c)}^{k} q_\phi(r \mid x) + \frac{1}{2}\sum_{r=k+1}^{N_t} q_\phi(r \mid x), \tag{40}$$

where the first sum is empty for $c = -1$. The first term is the posterior mass for which the gold-cover rank is at most k and the implied excess depth $k - r$ does not exceed capacity c. If $r > k$, the pool does not yet cover the gold set and order agreement is uninformative about capacity; the second term therefore assigns the maximum-entropy Bernoulli probability $1/2$. Thus s_c has no separately trained parameters. Given current reader posterior $p(c)$, the predictive stability probability is

$$\rho_k = \sum_c p(c) s_c(x,k). \tag{41}$$

Let $Z \in \{0,1\}$ be a hypothetical binary stability observation at candidate probe depth k, where $Z = 1$ denotes agreement and $Z = 0$ disagreement. Its predictive probability is ρ_k. The corresponding hypothetical capacity posteriors are

$$p(c \mid Z = 1, k) = \frac{p(c) s_c(x,k)}{\rho_k}, \tag{42}$$

$$p(c \mid Z = 0, k) = \frac{p(c)(1 - s_c(x,k))}{1 - \rho_k}. \tag{43}$$

For the current query, let $\widetilde{U}(K; p)$ denote the exposure score computed with capacity distribution p. Define $V(p) = \max_{K \in \{K_0(x),\ldots,N_t\}} \widetilde{U}(K; p)$ as the best available score. For candidate depth k, $\mathrm{EVI}(k)$ is the expected increase in this score after observing Z:

$$\mathrm{EVI}(k) = \rho_k V(p(\cdot \mid Z = 1, k)) + (1 - \rho_k) V(p(\cdot \mid Z = 0, k)) - V(p). \tag{44}$$

We query $K_p = \arg\max_k \mathrm{EVI}(k)$. This chooses a measurement for its ability to change the final decision, rather than because it is near a fixed depth.

C.4 Deterministic Same-Pool Perturbation

For a query with stable identifier d_x and global seed s, we compute

$$h_x = \mathrm{SHA256}(s : d_x) \tag{45}$$

and use the first eight digest bytes as the seed of a local pseudorandom permutation. If the resulting order is the identity, we apply a one-position cyclic rotation. The membership of the pool is unchanged. Consequently, the Dice signal in Eq. 11 measures presentation stability rather than retrieval coverage. Permutations are generated separately for each query and are reproducible across reruns.

C.5 Generalized Bayes Update

Let q be a candidate distribution over capacity states and q_x^* the distribution selected for query x. Here q is distinct from the coverage posterior q_ϕ. Let $\mathrm{KL}(q\|p_{m,t})$ denote the Kullback–Leibler divergence from the pre-query capacity distribution. For observed agreement $z_x \in [0,1]$, the generalized Bayes update minimizes

$$q_x^* = \arg\min_q \left\{\mathbb{E}_{c\sim q}\left[-z_x \log s_c(x,K_p) - (1-z_x)\log(1-s_c(x,K_p))\right] + \mathrm{KL}(q\|p_{m,t})\right\}. \quad (46)$$

The unique normalized solution is the posterior in Eq. 12 (Bissiri et al., 2016). This variational result establishes coherence relative to the chosen stability loss, not that same-pool agreement is a correctness event: identical wrong predictions can also yield $z_x = 1$. Accordingly, s_c is a working likelihood for presentation stability and its decision value is evaluated empirically. There is no temperature or manually chosen coefficient between the likelihood and prior.

C.6 End-to-End Algorithm

Algorithm 1: Capacity-adaptive schema exposure

1. **Source training:** learn $q_\phi(R \mid x)$ from three source domains; collect nested-prefix outcomes from source readers; fit transferable item difficulties and nuisance source-reader coordinates with Eq. 31 for initial probe design.
2. **One-time reader adaptation:** build synthetic target-schema tests; acquire informative nested-prefix outcomes; update the interval-censored boundaries with Eq. 32; transform Eqs. 34–35 into $p_{m,t}(c)$ and cache it.
3. **For each target query:** rank target intents, compute $q_\phi(R \mid x)$ and K_0; select K_p using Eq. 44; obtain canonical and permuted predictions; compute z_x; update the capacity posterior with Eq. 12; choose K^* using Eq. 13; make the final frozen-reader prediction with $_{K^*}(x)$.

D Data, Prompts, Readers, and Baselines

D.1 Dataset Statistics

Table 9: Dataset statistics in the released experiment files. Intent counts refer to the target schema after normalization.

Suite	Target	Intents	Train	Dev	Test	Total
Blend	ATIS	18	20,250	1,125	1,125	22,500
Blend	SNIPS	7	50,625	2,613	2,615	55,853
Blend	Banking77	77	36,390	2,009	2,021	40,420
Blend	CLINC150	147	54,896	2,889	2,977	60,762
Mix	ATIS	18	18,000	1,000	1,000	20,000
Mix	SNIPS	7	45,000	2,500	2,500	50,000
Mix	Banking77	77	32,340	1,796	1,796	35,932
Mix	CLINC150	147	48,824	2,639	2,639	54,102

Blend and Mix differ in how one-to-three single-intent utterances are composed into multi-intent examples. Every comparison uses identical records and rankings within a setting. "Overall" always means the unweighted macro average over the eight held-out target settings, or over the explicitly named reader subset; it never means a micro-average over examples.

D.2 Deployment-Query Availability

The evaluation uses a transductive deployment protocol. Before any target-reader calibration call, the complete target test split is available as the unlabeled batch $\mathcal{U}_t$. The frozen Stage 1 selector computes one required-rank posterior $q_\phi(r \mid x)$ per utterance. Their empirical distribution guides only which description-derived synthetic boundary is queried next during the one-time reader calibration. No

target label and no target-reader prediction on a real test utterance is available at this stage. For native CAPE, each utterance receives its two order-stability calls and final prediction after calibration is fixed. Consequently, all reported deployment results are label-free but transductive rather than schema-only.

D.3 FROZEN-READER PROMPT

All local and hosted readers receive provider-equivalent system and user messages. Intent IDs are local to the displayed pool, preventing memorized global label indices from influencing results. Outputs use JavaScript Object Notation (JSON).

> **System.** You are a precise multi-intent classifier. Identify the distinct semantic requests expressed in the utterance. Map each request to the single best matching intent from the provided intent descriptions. Include an intent only when it is directly supported by the utterance, and do not add related but unexpressed intents. Merge duplicate predictions. Select only integer IDs from the provided list. Return only a JSON array of integer IDs and no explanation.
>
> **User.** Dataset: {`dataset`}
> Allowed intent IDs, labels, and descriptions:
> `1. {intent name}: {description}`
> `...`
> Utterance: {`utterance`}
> Map every distinct semantic request in the utterance to its best-supported intent ID. Return one ID when the utterance expresses only one request. The JSON array must be non-empty and contain no more than three distinct IDs. Return only the JSON array.

The one-to-three output bound follows the construction of all evaluation datasets and is shared by every method. It is not a pool-selection parameter. Local-reader classification uses greedy generation (`do_sample=False`), at most 96 new tokens, and stops after a complete JSON array. Hosted readers use temperature zero and the same 96-token output limit. These settings are shared by Full, fixed-depth, Adaptive-k, LDAR, and CAPE. Three-view controls aggregate their prediction sets; the arbitration variant in Appendix P instead uses conditional confirmation.

D.4 NATURALIZATION PROMPT

> **System.** You convert one semantic request specification into one realistic user utterance clause for classifier calibration. Preserve the requested meaning, semantic roles, subject, object, tense, and temporal direction. Write from the first-person perspective of the user making the request. Introduce no additional request. Return one clause, not a list of alternatives. Do not emit an annotation, explanation, or intent label.
>
> **User.** Rewrite the semantic request below as one natural conversational clause.
> Semantic request: {`held-out intent description`}
> Return only a JSON object with this form: {`"clause": "..."`}.

Naturalization uses greedy decoding. A parser accepts the requested JSON, a single quoted string, or a labeled one-clause response. An invalid initial response is retried once with a stricter rewriting prompt.

D.5 READERS

Table 10 summarizes the evaluated readers, their roles, and their inference precision.

D.6 BASELINE DEFINITIONS

Full. All target intent descriptions are displayed in cosine-rank order.

Table 10: Reader roles and inference precision. "Source" means that the reader contributes source-domain item-response observations, not that its parameters are updated.

Reader	Size	Role	Precision
Qwen3	0.6B	Source	BF16
Qwen3	1.7B	Held-out size	BF16
Llama 3.2	1B	Held-out family	BF16
Llama 3.2	3B	Held-out family	BF16
Qwen3	4B	Source	BF16
Llama 3.1	8B	Held-out family	BF16
Qwen3	8B	Held-out size	BF16
Qwen3	14B	Source	4-bit
Qwen3	32B	Held-out size	4-bit
GPT-4o-mini	undisclosed	Hosted held-out	Provider
Gemini Flash-Lite	undisclosed	Hosted held-out	Provider

Fixed K. The first 5, 10, or 20 ranked intents are displayed; K is capped by schema cardinality. The historical result archive evaluated requested depths in separate greedy runs. Across the 54 requested-depth comparisons that collapse to Full (SNIPS $K = 10, 20$ and ATIS $K = 20$, over two suites and nine readers), the ranked pools and prompt-token counts are identical for every query. The separately decoded predictions agree on 99.53% of queries; their setting-macro fixed-minus-Full difference is -0.04 F1 and -0.04 Exact points, and the largest absolute setting Exact difference is 0.36 points. Local classification uses `do_sample=False`; small residual differences can nevertheless arise from separate low-precision GPU executions and batching. The released evaluator now copies an existing output for an identical (utterance, ordered pool, prompt) key, making duplicate effective policies bit-identical on replay. We keep the archived rows visible but collapse them to Full in policy diagnostics.

Adaptive-k. Following Taguchi et al. (2025), the cutoff is the largest adjacent similarity gap plus the method's prescribed downstream buffer. We use the authors' public implementation choices and do not tune the rule on target labels.

LDAR. We adapt the distraction-aware similarity-band policy of Shim et al. (2026) to intent-description ranking. For each reader and held-out target, the policy receives the sorted query–description cosine distribution and predicts Beta distributions over a lower quantile and band width. It uses 256-dimensional periodic embeddings, a two-layer four-head Transformer, and learned attention pooling. We optimize the policy by REINFORCE Williams (1992) with frozen-reader exact-set reward, Adam at 3×10^{-4}, batch size 32, and an exponential-moving-average baseline of 0.5. For the main comparisons, a deterministic, compute-bounded budget of 64 training and 32 development utterances per source domain is used; source-development macro exact match selects among at most 30 epochs, evaluated every five epochs with patience three. No target utterance or target label enters policy fitting or checkpoint selection. This protocol is source-domain matched but not source-example matched to CAPE: the coverage selector uses every source train/dev utterance in Table 9, while LDAR uses 192/96 distinct train/dev utterances in each three-source run. For example, when Blend CLINC150 is held out, the coverage selector uses 107,265/5,747 source train/dev utterances. Conversely, LDAR observes reader-specific source rewards for every deployed reader, including readers excluded from our source response-geometry fitting. The LDAR cap is a prespecified bound on expensive reader-specific policy optimization. Appendix N evaluates sensitivity to a fourfold increase in source training data on two held-out readers.

DSCP. We reproduce the DSCP prompt of Qin et al. (2025): divide the utterance into single-request clauses, solve each clause against the displayed schema, and combine duplicate-free intent IDs. We retain its native reasoning procedure in both controls: the native condition exposes the Full schema, while the controlled condition applies CAPE before the same DSCP steps. DSCP changes the reasoning protocol rather than estimating a query–reader-specific pool size, so this factorial tests whether adaptive exposure complements a distinct reasoning procedure.

Compute-matched controls. We rerun each exposure policy under three deterministic schema-order views: the canonical ranking and two non-identity permutations of the same pool. We then select the observed prediction set with the greatest total Sørensen–Dice agreement, breaking ties in canonical-then-permutation order. Thus these controls match CAPE in query-level reader calls without changing candidate membership or adapting exposure from the responses. The $K = 20$, Adaptive-k, and LDAR three-view controls are complete over all 72 local settings. The separate DSCP exposure factorial is reported in Table 21.

E FULL PERFORMANCE RESULTS

E.1 FULL LOCAL READER–TARGET MATRIX

Table 11 restores the reader–target granularity compressed by the two panels of Main Table 1. Every row averages only Blend and Mix; hosted readers are excluded.

Table 11: **Complete local reader–target performance matrix.** Each cell reports sample-F1/exact (%), averaged over Blend and Mix. The 36 rows cover nine local readers and four held-out target schemas. †Reader used to fit source response geometry; ‡4-bit loading. Every LDAR cell averages a complete Blend/Mix pair; no missing value is imputed.

Reader	Target (N)	Full	$K = 5$	$K = 10$	$K = 20$	Adaptive-k	LDAR	CAPE
Panel A: Source Qwen readers (used to fit response geometry)								
Qwen3-0.6B†	SNIPS (7)	61.39/22.12	66.40/29.28	61.30/21.94	61.30/21.94	62.16/20.81	67.63/29.16	75.63/44.96
	ATIS (18)	46.20/15.03	46.45/10.56	48.23/8.92	45.61/14.89	45.52/7.46	47.92/11.88	52.13/22.57
	Banking77 (77)	45.04/16.77	40.49/9.10	40.04/9.59	40.03/13.04	39.81/7.75	42.38/13.90	45.38/17.69
	CLINC150 (147)	54.59/20.52	56.63/21.79	53.62/18.51	53.48/23.39	54.06/16.30	55.34/19.40	58.39/27.28
Qwen3-4B†	SNIPS (7)	90.35/79.95	89.52/78.45	90.32/79.95	90.32/79.95	90.10/79.49	82.10/60.07	89.84/79.20
	ATIS (18)	83.27/64.97	72.73/48.38	80.87/60.79	83.09/64.82	80.82/60.56	77.99/55.77	78.11/57.37
	Banking77 (77)	64.58/42.43	60.27/36.21	65.03/41.75	66.05/43.29	63.31/39.22	65.97/43.06	66.19/42.82
	CLINC150 (147)	79.38/60.53	75.58/51.61	80.54/59.60	82.60/63.81	78.48/56.25	80.53/61.97	82.74/64.28
Qwen3-14B†,‡	SNIPS (7)	93.61/83.28	92.76/81.87	93.61/83.28	93.61/83.28	93.43/83.13	91.07/77.98	93.20/82.80
	ATIS (18)	84.45/59.32	75.98/49.08	83.15/57.19	84.45/59.32	83.37/58.83	82.34/56.21	79.81/55.10
	Banking77 (77)	73.10/49.56	65.51/40.33	69.73/44.50	71.67/46.75	68.94/43.63	73.36/49.45	71.64/47.67
	CLINC150 (147)	88.16/71.68	79.12/54.19	83.54/62.67	86.67/68.48	82.23/59.94	88.71/73.12	88.91/73.48
Panel B: Held-out Qwen sizes (unseen during response-geometry fitting)								
Qwen3-1.7B	SNIPS (7)	75.46/48.16	76.34/49.75	75.53/48.24	75.53/48.24	75.09/48.25	70.12/35.58	78.82/50.41
	ATIS (18)	62.26/29.88	59.39/27.10	62.82/28.93	62.33/29.97	61.84/28.76	59.66/27.45	60.65/28.51
	Banking77 (77)	44.55/21.83	51.06/24.90	49.86/22.95	47.79/20.14	48.53/23.38	44.77/19.75	52.58/26.00
	CLINC150 (147)	51.24/26.77	55.67/29.66	55.40/29.43	58.17/30.80	54.50/28.98	51.56/26.12	61.07/33.76
Qwen3-8B	SNIPS (7)	92.23/82.19	91.17/80.19	92.24/82.23	92.24/82.23	91.79/81.54	88.69/73.87	91.56/81.21
	ATIS (18)	82.43/63.57	73.10/47.88	81.41/61.18	82.41/63.52	80.54/60.02	81.37/60.93	76.93/54.11
	Banking77 (77)	63.45/39.59	61.45/36.69	65.61/41.81	65.41/41.62	63.73/39.31	64.43/40.13	66.73/42.72
	CLINC150 (147)	78.86/59.52	76.47/52.53	80.50/59.93	81.88/63.33	79.07/57.13	81.19/62.89	83.22/64.80
Qwen3-32B‡	SNIPS (7)	93.76/85.91	93.34/84.53	93.76/85.91	93.76/85.91	93.62/85.68	92.01/80.97	93.89/85.67
	ATIS (18)	82.69/55.66	73.17/43.25	81.11/52.43	82.69/55.66	80.72/53.13	80.03/52.08	78.07/52.43
	Banking77 (77)	73.08/51.12	65.58/39.98	70.19/46.24	72.26/49.35	69.24/44.42	73.12/51.22	72.74/49.62
	CLINC150 (147)	89.38/75.72	80.06/54.58	85.08/64.79	87.86/71.13	83.52/61.67	89.47/75.56	90.07/76.10
Panel C: Held-out Llama readers (unseen model family)								
Llama-3.2-1B	SNIPS (7)	59.35/4.73	64.51/5.97	59.27/4.67	59.27/4.67	59.74/4.61	64.13/24.76	76.86/41.85
	ATIS (18)	36.65/0.66	46.43/1.55	41.51/0.71	36.69/0.61	41.29/0.89	51.12/22.42	54.56/13.23
	Banking77 (77)	7.22/0.29	41.74/1.72	36.50/1.20	25.56/0.13	35.77/1.23	47.78/16.98	44.29/13.59
	CLINC150 (147)	26.12/4.35	50.10/2.15	44.70/2.01	41.59/2.60	46.16/2.04	51.74/7.23	50.55/17.82
Llama-3.2-3B	SNIPS (7)	73.23/37.68	73.25/40.11	73.17/37.56	73.17/37.56	73.53/39.43	63.21/24.50	78.78/50.02
	ATIS (18)	55.63/12.83	57.62/21.78	56.55/14.13	55.54/12.65	57.52/19.41	59.15/20.41	60.78/26.25
	Banking77 (77)	46.88/6.43	51.91/17.00	50.75/11.05	50.58/15.37	51.10/14.12	49.30/10.43	54.81/21.12
	CLINC150 (147)	56.27/14.09	60.94/24.08	62.99/24.78	63.94/30.59	62.69/26.30	63.55/31.19	65.02/32.65
Llama-3.1-8B	SNIPS (7)	80.10/49.15	78.49/46.86	80.13/49.15	80.13/49.15	79.23/47.90	62.66/24.39	77.59/49.08
	ATIS (18)	63.81/18.11	61.11/19.24	66.90/22.93	63.77/18.06	64.47/21.17	62.66/20.81	64.30/27.53
	Banking77 (77)	51.76/10.06	55.34/19.13	57.27/16.73	55.18/13.99	56.44/18.51	53.30/11.96	58.21/23.75
	CLINC150 (147)	65.60/25.96	66.59/32.12	70.87/34.52	68.99/29.94	69.61/35.11	68.34/29.89	71.77/39.41

Table 12: **Complete hosted-API results.** Each method cell is Blend/Mix macro-averaged sample-F1/exact (%). $\bar{K}$ is the mean exposure selected by CAPE, also macro-averaged over Blend and Mix. The LDAR column is populated from the independently completed hosted-reader setting artifact and aggregated identically.

Reader	Target	Full	$K = 5$	$K = 10$	$K = 20$	Adaptive-k	LDAR	CAPE	$\bar{K}$
GPT-4o-mini	SNIPS	93.69/86.51	93.13/84.81	93.69/86.52	93.69/86.52	93.64/86.03	93.24/84.96	93.50/85.81	5.57
	ATIS	83.60/63.23	75.03/47.79	82.43/59.87	83.48/62.97	82.13/60.16	84.01/64.23	78.73/55.23	5.75
	Banking77	74.98/53.67	65.73/39.86	70.72/47.04	72.92/50.76	69.38/44.80	75.29/54.11	75.18/53.78	13.63
	CLINC150	91.29/79.93	80.29/54.44	85.57/65.53	88.59/72.49	83.90/61.71	91.41/80.14	92.22/80.83	59.35
Gemini-3.1-Flash-Lite	SNIPS	95.93/91.20	94.88/88.46	95.93/91.20	95.93/91.20	95.80/90.99	94.27/86.68	95.08/89.12	5.63
	ATIS	86.80/67.61	76.36/49.32	85.31/63.80	86.80/67.61	84.72/63.98	86.62/67.49	81.95/59.54	5.86
	Banking77	80.69/62.90	68.62/42.97	74.53/51.28	78.07/57.39	73.92/49.81	80.76/62.85	81.05/63.19	15.89
	CLINC150	95.62/89.89	81.95/56.10	87.70/68.37	91.25/77.23	85.77/64.36	95.56/89.69	96.25/90.65	59.25

E.2 TARGET-WISE HOSTED-API RESULTS

The hosted-reader matrix is complete and is reported below. Each row is one hosted reader–target pair, with Blend and Mix macro-averaged using equal weight. Main Table 4 further aggregates these target-level results over either all targets or the two larger schemas.

F EXPOSURE BEHAVIOR

Table 13 summarizes target-level selection behavior. Figure 5 resolves the same quantity by local reader and target, distinguishing a reader-blind retrieval rule from policies that can adapt to the deployed reader.

Table 13: **Schema exposure and gold coverage.** Values average nine local readers and Blend/Mix. $\bar{K}$ is followed by its schema fraction; Cov. is complete gold-set inclusion, evaluated post hoc. Overall is the 72-setting macro.

	Adaptive-k		LDAR		CAPE	
Target (N)	$\bar{K}$ (% schema)	Cov.	$\bar{K}$ (% schema)	Cov.	$\bar{K}$ (% schema)	Cov.
SNIPS (7)	6.54 (93.4%)	99.0%	2.38 (34.0%)	54.0%	4.02 (57.5%)	81.6%
ATIS (18)	10.62 (59.0%)	90.5%	7.79 (43.3%)	73.1%	4.82 (26.8%)	67.2%
Banking77 (77)	16.78 (21.8%)	67.9%	35.45 (46.0%)	82.3%	8.53 (11.1%)	61.2%
CLINC150 (147)	13.32 (9.1%)	68.2%	62.22 (42.3%)	87.3%	27.70 (18.8%)	78.0%
Overall	**11.82**	**81.4%**	**26.96**	**74.2%**	**11.27**	**72.0%**

G SOURCE-READER ROBUSTNESS

Main Table 1 evaluates three unseen Qwen sizes and three unseen Llama readers, while Main Table 4 evaluates two hosted APIs. Here, we examine whether the source response model depends disproportionately on one of its three source readers. Each leave-one-source-reader-out (LOSO) run removes one reader's response observations and refits only the response model. The coverage selector, test readers, target examples, prompts, and evaluation matrix remain fixed.

Across the three omissions, overall F1 changes by -0.07 to $+0.01$ points and Exact by -0.32 to -0.14 points. Mean exposure ranges from 10.51 to 10.68. These small changes suggest that no single source reader dominates the learned response model. The control nevertheless remains limited to the reader families evaluated in the main matrix.

H TRAINING-SEED STABILITY

We retrain Stage 1 with three random seeds while holding the data splits, reader panel, test-time calibration protocol, and evaluation matrix fixed.

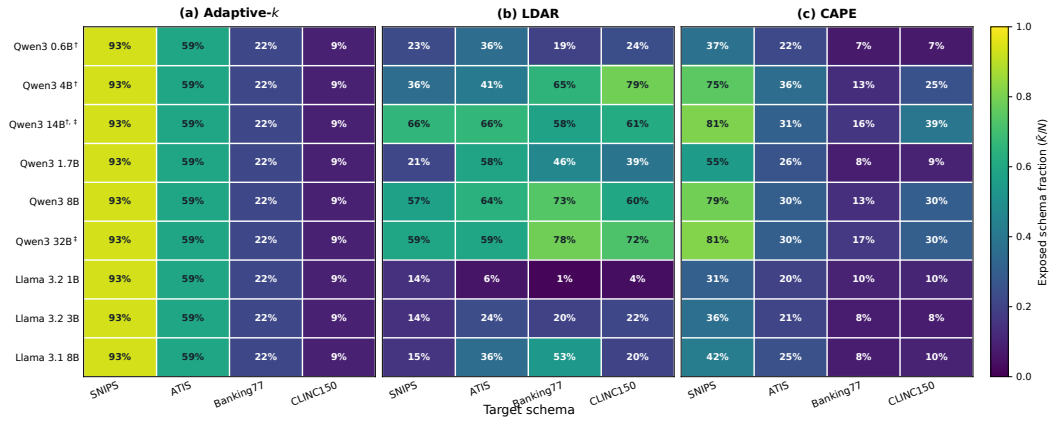


Figure 5: **Realized schema exposure by reader and target.** Each cell prints the schema-normalized exposure $\bar{K}/N$, computed after macro-averaging $\bar{K}$ over Blend and Mix; color encodes the same fraction on a shared scale. Panels show Adaptive-k, LDAR, and CAPE, respectively. Adaptive-k is reader-blind, whereas LDAR and CAPE can vary with the deployed reader. All reader–target cells contain complete Blend/Mix pairs. †Readers used to fit CAPE's source response geometry; ‡4-bit loading.

Table 14: **Leave-one-source-reader-out response-geometry control.** F1, Exact, and mean exposure are macro-averaged over all 72 local reader–target–suite settings. ΔExact is relative to the geometry fit with all three source readers. Only the source-reader observation panel changes between rows.

Source readers	Omitted reader	Scope	F1	Exact	$\bar{K}$	ΔExact
Qwen 0.6/4/14B	–	All 72	70.72	44.91	11.27	–
Qwen 4/14B	Qwen3-0.6B	All 72	70.65	44.71	10.55	−0.20
Qwen 0.6/14B	Qwen3-4B	All 72	70.73	44.77	10.51	−0.14
Qwen 0.6/4B	Qwen3-14B	All 72	70.70	44.59	10.68	−0.32

Table 15: **Stage 1 random-seed stability.** Each seed retrains the coverage selector and item-response model from the same source-only splits and evaluates the complete 72-setting local matrix. Large-schema Exact is the 36-setting Banking77/CLINC150 macro-average. The final row reports the mean and sample standard deviation across the three seeds.

Seed	F1	Exact	Large-schema Exact	$\bar{K}$
3407	70.72	44.91	39.70	11.27
3408	71.04	44.60	38.95	10.91
3409	70.51	45.42	40.45	11.63
Mean ± std	70.76 ± 0.27	44.98 ± 0.41	39.70 ± 0.75	11.27 ± 0.36

Table 15 reports variation from Stage 1 training while holding the deployed-reader calibration and evaluation protocol fixed.

I MECHANISM AND SENSITIVITY CONTROLS

We prioritize controls that test why the proposed construction is needed over a broad numerical hyperparameter sweep. Every comparison below keeps the frozen reader, cosine ranking, prompt, target examples, and evaluation scope fixed. It changes one source of evidence or one inference assumption.

I.1 Main-Component Ablation Definitions

Table 3 uses four prespecified decision paths and one reader-posterior intervention. All five rows share the same Stage 1 required-rank posterior $q_\phi(r \mid x)$, anchor $K_0(x) = \arg\max_r q_\phi(r \mid x)$, immutable cosine ranking, frozen reader, and final classification prompt. Only the reader- and query-level evidence paths differ. To state these paths precisely, define the decision functional

$$\mathcal{D}_x[p] = \operatorname*{arg\,max}_{K \in \{K_0(x), \ldots, N_t\}} \sum_{r=1}^{K} q_\phi(r \mid x) \sum_{c \geq K-r} p(c), \tag{47}$$

and the normalized query-evidence update

$$\mathcal{B}_{x,K_p,z}[p](c) = \frac{p(c) s_c(x, K_p)^z [1 - s_c(x, K_p)]^{1-z}}{\sum_{c'} p(c') s_{c'}(x, K_p)^z [1 - s_{c'}(x, K_p)]^{1-z}}. \tag{48}$$

Here $p^{\mathrm{RA}}_{m,t}(c)$ denotes the reader posterior inferred from the one-time target-schema self-tests. The query-only control instead starts from the finite-support maximum-entropy prior

$$p^{\mathrm{ME}}_t(c) = \frac{1}{N_t + 1}, \qquad c \in \{-1, 0, \ldots, N_t - 1\}. \tag{49}$$

This prior depends only on schema size and contains no reader-specific self-test evidence.

The four decision paths are therefore implemented as follows:

- **Stage 1 only:** $K^{\mathrm{S1}}(x) = K_0(x)$. It performs neither schema self-tests nor query agreement probes.
- **+ Reader adaptation:** $K^{\mathrm{R}}(x, m, t) = \mathcal{D}_x[p^{\mathrm{RA}}_{m,t}]$. It uses the one-time schema self-tests but performs no query probe and no update in Eq. 48.
- **+ Query adaptation:** choose K_p by the same EVI rule as the full method, but compute that EVI under p^{ME}_t. After observing same-pool agreement z_x, set

  $$p^{\mathrm{Q}}_{x,t} = \mathcal{B}_{x,K_p,z_x}[p^{\mathrm{ME}}_t], \qquad K^{\mathrm{Q}}(x, t) = \mathcal{D}_x[p^{\mathrm{Q}}_{x,t}]. \tag{50}$$

  It uses no target-schema self-test outcome or other reader-level evidence.
- **Full Cape:** choose K_p by EVI under $p^{\mathrm{RA}}_{m,t}$, set $p^{\mathrm{Full}}_{x,m,t} = \mathcal{B}_{x,K_p,z_x}[p^{\mathrm{RA}}_{m,t}]$, and return $K^{\mathrm{Full}}(x, m, t) = \mathcal{D}_x[p^{\mathrm{Full}}_{x,m,t}]$.

Thus "query adaptation only" does not silently inherit the deployed reader's calibration posterior, and "reader adaptation only" does not spend or use query-level agreement calls.

Performance improves steadily with the calibration budget, with diminishing returns from 32 to 64 reader calls. Single-intent calibration exposes substantially more intents yet reduces Exact, showing that multi-intent probes capture set-level fragility rather than merely encouraging broader exposure. The empirical stable-depth control is more conservative and performs worst, supporting retention of posterior uncertainty rather than collapsing each self-test to a hard depth. Fixed-depth probing and the raw capacity survival produce smaller degradations, indicating that adaptive acquisition and the shape constraint provide complementary refinements rather than solely driving the gain. The allocation control holds the synthetic candidate set, 64-call budget, initial design, posterior estimator, query adaptation, and final decision rule fixed. Its source-derived reference distribution maps schemas by linear interpolation of normalized schema-size rank, with exact ties resolved by alphabetical target name; this rule is fixed without target outcomes. Relative to that control, deployment-batch allocation improves overall Exact by 0.55 points and large-schema Exact by 0.63 points, while reducing mean exposure by 0.16 intents. Both allocation policies are evaluated on 28,800 matched reader–query cases from the fixed diagnostic subset.

Source-informed initialization. We replace Fisher-ranked initial probe selection with random selection from the same candidate bank, keeping the number of initial probes, anchor evaluations, and total 64-call calibration budget unchanged. Subsequent adaptive allocation, capacity estimation, query adaptation, and final selection follow the same procedure. Only the initial probe-selection seed varies across three runs; trained components remain fixed. Random initialization achieves 44.43 ± 0.14 overall Exact and 38.97 ± 0.15 large-schema Exact, compared with 44.91 and 39.70 under Fisher-ranked initialization.

Table 16: **Mechanism and sensitivity controls.** Panel A uses the complete test splits for all 72 local reader–target–suite settings. Panel B independently evaluates both allocation policies, including the default, on 28,800 matched reader–query cases from the fixed diagnostic subset. Overall Exact is macro-averaged over all 72 settings, and Large-schema Exact over the 36 Banking77 and CLINC150 settings. Rows labeled "default" provide the matched CAPE reference for each question. Random initialization reports mean ± standard deviation across three probe-selection seeds with trained components fixed.

Question	Variant	Overall Exact	Large-schema Exact	Mean K
Panel A: Full-test mechanism controls				
How much calibration is needed?	16 reader calls	43.85	38.62	10.14
	32 reader calls	44.53	39.38	10.82
	64 reader calls (default)	44.91	39.70	11.27
Are multi-intent self-tests useful?	Single-intent probes only	43.92	38.15	14.38
	One-to-three intents (default)	44.91	39.70	11.27
Does retaining calibration uncertainty matter?	Hard empirical stable depth	43.17	37.49	8.52
	Interval-censored capacity posterior (default)	44.91	39.70	11.27
Is adaptive probe selection needed?	Fixed at K_{reader}	44.68	39.21	11.05
	EVI-selected K_p (default)	44.91	39.70	11.27
Is the shape constraint consequential?	Raw capacity survival	44.75	39.54	11.33
	Isotonic capacity survival (default)	44.91	39.70	11.27
Does source-informed initialization matter?	Random initial probes, 3 seeds	44.43 ± 0.14	38.97 ± 0.15	11.48 ± 0.12
	Fisher-ranked initial probes (default)	44.91	39.70	11.27
Panel B: Deployment-batch allocation audit (28,800 matched cases)				
Does deployment-aware allocation help?	Source-reference allocation	44.62	39.21	11.58
	Deployment-batch allocation (default)	45.17	39.84	11.42

I.2 Simple Calibrated Policy Controls

We test whether the same synthetic self-test budget supports a simpler policy that can replace the full posterior decision. All rows share the frozen ranker, source-trained coverage selector, synthetic candidate bank, classification prompt, and a maximum budget of 64 target-reader response calls. Target labels are used only after every policy and prediction has been fixed.

The *synthetic-selected fixed K* control evaluates the prespecified effective fixed-depth candidates on synthetic queries, chooses the depth with highest synthetic Exact for each reader–target pair, and uses that one depth for every real query in the pair; ties select the smaller depth. Its three-view version preserves this pool membership and applies the same canonical-plus-two-permutations aggregation used by the matched inference controls.

The coverage-threshold control selects a query-dependent depth from the same source-trained coverage posterior:

$$K_\tau(x) = \min\{K \in \{1, \dots, N_t\} : C_\phi(K \mid x) \geq \tau\}. \tag{51}$$

Let $\mathcal{B}_t$ denote the synthetic candidate bank. We collect the distinct breakpoints of the coverage cumulative distribution function (CDF) and select eight threshold candidates at evenly spaced empirical quantiles:

$$\begin{aligned} \mathcal{V}_t &= \text{unique}\{C_\phi(K \mid x_b) : x_b \in \mathcal{B}_t,\ K = 1, \dots, N_t\}, \\ \mathcal{T}_t &= \left\{ Q_{\mathcal{V}_t}\left(\frac{j - \frac{1}{2}}{8}\right) : j = 1, \dots, 8 \right\}, \end{aligned} \tag{52}$$

where $Q_{\mathcal{V}_t}$ is the empirical quantile function of the distinct breakpoint values. Candidate thresholds are determined without calibration responses or target labels.

For calibration budgets $B \in \{16, 32, 64\}$, the eight target-specific threshold candidates above remain fixed. We use nested subsets of the eight highest-ranked probes from the source Fisher design, $\mathcal{P}_2 \subset \mathcal{P}_4 \subset \mathcal{P}_8$. At each budget, all eight thresholds are evaluated on the same $n_B = B/8$ probe identities. Each call classifies one probe using the prefix at $K_\tau(x_b)$, giving $8 \times 2 = 16$, $8 \times 4 = 32$, or $8 \times 8 = 64$ synthetic classification calls per reader–schema pair. We score each threshold by

$$\widehat{A}_B(\tau) = \frac{1}{n_B} \sum_{b \in \mathcal{P}_{n_B}} \mathbf{1}\left[\widehat{Y}_{K_\tau(x_b)} = Y_b^{\text{syn}}\right]. \tag{53}$$

For each reader–schema pair and budget, we select the threshold with the highest synthetic Exact, breaking ties by smaller mean exposure on the evaluated probes and then smaller τ. The selected threshold is fixed for all real queries. Each query uses canonical and two permuted orders of its selected pool with the same three-view aggregation. This control adapts depth to query coverage without a capacity posterior or a query-level capacity update. The budget B counts only synthetic classification calls; probe naturalization is accounted for separately in Appendix K.

The midpoint capacity control instead maps every interval-censored synthetic boundary $[\ell_b, u_b]$ to

$$\widehat{B}_b = \left\lfloor \frac{\ell_b + u_b}{2} \right\rfloor, \qquad \widehat{c}_b = \widehat{B}_b - R_b, \tag{54}$$

then keeps risk-set handling, isotonic projection, query updating, and final selection identical to CAPE. Thus it preserves variation across probes but removes uncertainty within each probe's unresolved interval. EVI and the query-probe depth are recomputed under the midpoint posterior; evidence observed at a different depth is not reused.

Three-view aggregation raises the synthetic-selected fixed policy by 1.66 Exact points. Under matched calibration budgets and three query calls, CAPE improves over this policy by 1.73 points [1.42, 2.04] overall and 1.85 [1.28, 2.42] on held-out readers. The stronger coverage-threshold control reaches 69.95 F1 and 44.05 Exact with mean exposure 12.80. CAPE improves Exact by 0.86 points [0.61, 1.15] overall and 0.90 [0.32, 1.48] on held-out readers, while reducing mean exposure to 11.27. These comparisons support the combined retention modeling and query update beyond both reader-level fixed-depth calibration and query-dependent coverage thresholding.

Retaining interval uncertainty yields smaller additional gains over midpoint capacity: 0.46 Exact points [0.17, 0.74] overall and 0.58 [0.05, 1.11] on held-out readers. Overall F1 also increases by 0.36 points, while mean exposure decreases by 0.18 intents. The held-out interval narrowly excludes zero, so this comparison supports a modest refinement rather than the primary source of improvement.

Table 17: **Can simpler calibrated policies replace CAPE?** All policies share the synthetic candidate bank and a maximum budget of 64 synthetic classification calls per reader–schema pair. Panel A reports macro averages over 72 local settings; Calls denotes reader calls per query. Three-call fixed-K, coverage-threshold, and CAPE basic policies use three-view aggregation. Panel B reports full CAPE minus the three-call controls in Panel A. Panel C compares the simplified capacity policy with coverage thresholding and full CAPE; all three use three query calls, and mean K averages all 72 settings. Panel D reports paired differences from Panel C. F1 and Exact are percentages; differences are in percentage points with 95% paired query-bootstrap CIs. All nine and Held-out six cover 72 and 48 settings, respectively.

A. Performance and exposure

Policy	Calls	F1	Exact	Mean K
Synthetic-selected fixed K	1	68.45	41.52	15.34
Synthetic-selected fixed K	3	68.83	43.18	15.34
Coverage threshold $K_\tau(x)$	3	69.95	44.05	12.80
Midpoint capacity	3	70.36	44.45	11.45
CAPE interval posterior	3	**70.72**	**44.91**	**11.27**

B. CAPE advantage: ΔExact [95% CI] (pp)

Three-call policy	All nine	Held-out six
Synthetic-selected fixed K	$+1.73$ $[+1.42, +2.04]$	$+1.85$ $[+1.28, +2.42]$
Coverage threshold $K_\tau(x)$	$+0.86$ $[+0.61, +1.15]$	$+0.90$ $[+0.32, +1.48]$
Midpoint capacity	$+0.46$ $[+0.17, +0.74]$	$+0.58$ $[+0.05, +1.11]$

C. Simplified capacity policy: three query calls

Method	All-nine Exact	Held-out-six Exact	Mean K
K_τ + 3-view	44.05	40.83	12.80
CAPE basic	44.42	41.21	**10.85**
CAPE	**44.91**	**41.73**	11.27

D. Paired differences: ΔExact [95% CI] (pp)

Comparison	All nine	Held-out six
CAPE basic − (K_τ + 3-view)	$+0.37$ $[+0.09, +0.66]$	$+0.38$ $[-0.04, +0.81]$
CAPE − CAPE basic	$+0.49$ $[+0.22, +0.78]$	$+0.52$ $[+0.14, +0.95]$

Simplified capacity policy. CAPE basic jointly simplifies calibration and inference while retaining interval-based capacity estimation. It uses the same frozen ranker, coverage selector, synthetic candidate bank, classification prompt, and decoding settings. Probes follow a fixed seeded order instead of Fisher ranking, and a maximum of 64 synthetic classification calls is allocated by round-robin bisection of unresolved boundary intervals. We re-estimate capacity from the observations collected by this schedule, retaining the interval posterior, risk-set handling, and isotonic projection. The resulting capacity survival is used in the pre-query decision rule of Eq. 10 to select a query-dependent depth K_{reader}. At this fixed pool, we aggregate predictions from the canonical order and two permutations using the same Dice-based rule as the three-view controls. The policy uses neither IRT/Fisher probe design nor EVI calibration allocation or query-probe selection, and performs no query-level generalized-Bayes update. Unlike the midpoint control, it retains within-probe interval uncertainty and changes both calibration and inference.

Table 17, Panels C–D, reports the results. CAPE basic reaches 44.42 Exact overall and 41.21 on held-out readers, with mean exposure 10.85. It exceeds K_τ with three-view aggregation by 0.37 Exact points [0.09, 0.66] overall while using a smaller mean pool (10.85 vs. 12.80). The held-out difference is 0.38 points $[-0.04, 0.81]$; its interval includes zero. Full CAPE improves on the basic variant by 0.49 points [0.22, 0.78] overall and 0.52 [0.14, 0.95] on held-out readers. These results support a

useful simplified capacity policy and further gains from the full procedure. Because this comparison changes both calibration and final prediction aggregation, it does not isolate the contribution of any one component.

We compute 95% percentile CIs using 10,000 paired query-bootstrap replicates over the complete test splits. Within each target–suite split, we resample query IDs and carry all corresponding reader and policy outcomes together. Each replicate recomputes the equally weighted setting-macro difference over all 72 settings or the 48 held-out-reader settings. The held-out scope excludes Qwen3-0.6B, Qwen3-4B, and Qwen3-14B, which contribute source response-geometry observations. The intervals condition on fixed trained models, readers, and schema calibration.

Calibration-budget sensitivity. We compare CAPE and the coverage-threshold policy at calibration budgets of 16, 32, and 64 calls, keeping the frozen models, synthetic candidate bank, and three-call query budget fixed. Figure 3 reports macro Exact over all 72 local settings and the 48 held-out-reader settings. The matched-budget gains are 0.77, 0.82, and 0.86 points overall, and 0.74, 0.81, and 0.90 on held-out readers. With 32 calibration calls, CAPE reaches 44.53 Exact overall and 41.42 on held-out readers, exceeding the 64-call control's 44.05 and 40.83, respectively. These results support more effective use of limited calibration calls. Total-token accounting is reported separately in Appendix K.

Analysis by relative exposure. We compare the frozen $B = 64$ runs with seed 3407 over all 72 reader–suite–target settings. Both policies share the frozen ranking, coverage selector, calibration candidate bank, classification prompt, and decoding settings, with three reader calls per query. Each policy's calibration results and final predictions are held fixed during this audit. We group matched queries by whether CAPE's final depth K_C is smaller than, equal to, or larger than K_τ.

For setting s with n_s queries, assign each query weight $w_{si} = 1/(72n_s)$. For policy $h \in \{C, \tau\}$, define exact correctness and complete gold coverage as

$$E^h_{si} = \mathbf{1}[\widehat{Y}^h_{si} = Y_{si}], \qquad G^h_{si} = \mathbf{1}[Y_{si} \subseteq P^h_{si}], \tag{55}$$

where $\widehat{Y}^h_{si}$ is the final prediction and P^h_{si} the selected pool. Let $\mathcal{Q}_g$ contain the query records in group g. Its fraction, within-group Exact difference, and contribution to the overall difference are

$$\begin{aligned} p_g &= \sum_{(s,i)\in\mathcal{Q}_g} w_{si}, \\ \Delta\mathrm{Exact}_g &= \frac{100}{p_g} \sum_{(s,i)\in\mathcal{Q}_g} w_{si}(E^C_{si} - E^\tau_{si}), \\ c_g &= p_g\, \Delta\mathrm{Exact}_g. \end{aligned} \tag{56}$$

Thus $\sum_g c_g$ equals the 72-setting macro Exact difference. Replacing E with G gives the within-group coverage difference. Wrong-to-right and right-to-wrong rates use the same within-group weights; their difference equals $\Delta\mathrm{Exact}_g$. Gold labels are used only for this post-hoc audit. Intervals follow the paired query-bootstrap protocol above and condition on the fixed models and calibration.

Table 18 shows that larger-pool cases contribute $+1.07$ Exact points, with a 95% CI of $[0.80, 1.37]$, offsetting the -0.27-point contribution from equal-pool cases. The smaller-pool contribution is $+0.07$ points, with a 95% CI of [0.02, 0.12]. Overall, CAPE gains 0.86 Exact points while gold coverage decreases by 1.48 points. Equal depths give identical pools under the shared ranking, but final predictions can differ because the coverage-threshold policy uses three-view aggregation and CAPE uses two evidence calls followed by a final prediction. This analysis decomposes the difference between the complete policies; it does not isolate a causal effect of changing exposure alone.

I.3 Empirical Reader-Calibration Control

The empirical baseline uses exactly the same source-designed synthetic specifications, schema prefixes, reader responses, and 64-call calibration budget as CAPE. It differs only in how those target-reader outcomes are summarized: it collapses each interval-censored boundary posterior to one

Table 18: **CAPE versus coverage thresholding by selected exposure.** Results use $B = 64$, seed 3407, and all 72 local settings. Panel A reports selection fractions, within-group differences, and contributions to the overall Exact difference. Intervals below Exact differences and contributions are paired-bootstrap 95% CIs. Cov. measures observed complete gold-set inclusion. Panel B reports 95% CIs for within-group prediction-transition rates. Differences in displayed totals reflect rounding.

A. Observed differences and net contributions

Relative exposure	Selection (%)	ΔExact (pp)	Net contribution (pp)	ΔCov. (pp)
Smaller: $K_C < K_\tau$	9.56	$+0.72$ $[+0.21, +1.28]$	$+0.07$ $[+0.02, +0.12]$	-24.83
Same: $K_C = K_\tau$	47.15	-0.58 $[-0.86, -0.31]$	-0.27 $[-0.41, -0.15]$	0.00
Larger: $K_C > K_\tau$	43.29	$+2.46$ $[+1.84, +3.16]$	$+1.07$ $[+0.80, +1.37]$	$+2.06$
Overall	**100.00**	$\mathbf{+0.86}$ $[+0.61, +1.15]$	$\mathbf{+0.86}$ $[+0.61, +1.15]$	-1.48

B. Prediction transitions: 95% CIs (%)

Relative exposure	Wrong → right	Right → wrong
Smaller: $K_C < K_\tau$	$[5.82, 7.41]$	$[5.11, 6.63]$
Same: $K_C = K_\tau$	$[0.02, 0.17]$	$[0.46, 0.88]$
Larger: $K_C > K_\tau$	$[8.73, 10.32]$	$[6.34, 7.91]$
Overall	$[4.32, 5.14]$	$[3.61, 4.38]$

hard observed stable depth instead of retaining unresolved boundary mass. For synthetic probe b, let c_b be the largest observed value of $k - R_b$ up to which every queried nested endpoint remains exact; set $c_b = -1$ when the anchor observation is inexact. Unqueried larger depths are not credited. Let n_{probe} be the number of self-tests summarized by this control. The empirical capacity distribution is

$$p^{\text{emp}}_{m,t}(c) = \frac{1}{n_{\text{probe}}} \sum_{b=1}^{n_{\text{probe}}} [c_b = c]. \tag{57}$$

This deliberately stricter pathwise construction is an empirical control, not the endpoint label used to train Eq. 28. Its corresponding capacity survival is

$$S^{\text{emp}}_{m,t}(d) = \sum_{c} p^{\text{emp}}_{m,t}(c)[c \geq d]. \tag{58}$$

We substitute $S^{\text{emp}}_{m,t}$ for $S_{m,t}$ in Eq. 21 and leave query updating and final discrete selection unchanged. This control therefore asks whether a hard empirical summary of the same self-tests is sufficient, rather than comparing against no reader adaptation.

I.4 Fixed versus EVI-Selected Query Probes

The fixed-probe control spends the same two query-level evidence calls and one final prediction call as CAPE. It probes the canonical and deterministically permuted orders at the pre-query decision $K_{\text{reader}}(x, m, t)$, applies the same generalized Bayes update, and then recomputes the final depth. The EVI variant differs only by choosing

$$K_p = \arg\max_k \text{EVI}(k) \tag{59}$$

before those two calls. Thus any difference isolates adaptive acquisition of query evidence rather than call count, pool perturbation, or posterior updating.

I.5 Reader-Posterior Swap

The posterior-swap row in Table 3 directly intervenes on reader conditioning. Within each target schema and Blend/Mix condition, a SHA256-seeded ordering of the nine readers defines a one-cycle derangement. No reader retains its own pre-query posterior, and every posterior is used exactly once in that setting. We rerun query adaptation and final selection while holding the target query, cosine ranking, Stage 1 posterior, prompt, and tested reader fixed. Hence the only changed object is the schema-conditioned capacity posterior,

$$p_{m,t}(c) \longleftarrow p_{m',t}(c), \qquad m' \neq m. \tag{60}$$

Across all 72 settings, the shuffled posterior obtains 69.95 F1 and 43.40 Exact, compared with 70.72 and 44.91 for full CAPE. Its mean exposure is smaller (10.37 versus 11.27), yet its gold-set coverage is slightly higher (72.97 versus 71.99). The degradation despite higher coverage supports matching the capacity posterior to the tested reader. Higher gold coverage alone does not explain the performance difference.

I.6 Naturalization Artifact Controls

All naturalization controls use the same target intent names, held-out description specifications, synthetic gold sets, probe scheduling rule, and response budget. They differ only in the text shown as a synthetic query:

- **Same-reader naturalization** asks the tested reader to rewrite each specification, as in the current method.
- **Independent naturalization** uses one naturalizer fixed before evaluation, Qwen2.5-7B-Instruct (Qwen Team, 2025a), and distinct from every tested reader. It is not selected using target utterances or target labels.
- **Raw description** presents the held-out specification without generative rewriting.

For each condition, we report sample-F1, exact match, and mean K as macros over all 72 local reader–target–suite settings. Capacity rank correlation is Spearman's ρ against the default condition, computed across the nine readers independently for each target and then macro-averaged across the four targets. All three conditions contain the complete 72 local reader–target–suite settings, for 216 matched condition–setting results. The independent naturalizer preserves the default capacity ordering with $\rho = 0.946$ and obtains 70.70 F1, 44.76 Exact, and $\bar{K} = 11.04$, compared with 70.72, 44.91, and 11.27 under same-reader naturalization. Raw descriptions preserve the ordering with $\rho = 0.975$ and obtain 70.77 F1, 44.91 Exact, and $\bar{K} = 10.75$. Across individual targets, the correlations are 0.867–0.983 for independent naturalization and 0.967–0.983 for raw descriptions. Performance and reader ordering therefore persist when the tested reader does not write its own calibration text, arguing against self-wording consistency as the source of the gain.

I.7 Shape and Numerical Checks

Table 16 tests the operational consequence of the shape constraint by running the complete decision pipeline with raw versus isotonic capacity survival. Without projection, the method obtains 44.75 overall Exact, 39.54 large-schema Exact, and $\bar{K} = 11.33$, compared with 44.91, 39.70, and 11.27 under the default projection. The small but consistent differences show that isotonic projection is a validity-preserving refinement rather than the primary source of the method's gain.

J Matched-Precision Control

Because Qwen3-14B and 32B use 4-bit loading, model scale and numerical precision are otherwise confounded. We isolate this issue with one reader, Qwen3-8B, evaluated in BF16 and 4-bit precision on the two larger schemas. The relevant claim is not that quantization is harmless, but that the capacity–schema trend is qualitatively stable under matched reader weights.

Under 4-bit loading, CAPE retains Exact advantages over Full of $+2.65$ and $+6.54$ points on Banking77 and CLINC150, respectively, and advantages over Adaptive-k of $+4.02$ and $+8.12$ points.

Table 19: **Matched-precision control on the two larger schemas.** Method cells are sample-F1/exact (%), macro-averaged over Blend and Mix. K_{CAPE} is the mean number of intents selected by CAPE. Both precision conditions use the same Qwen3-8B reader weights.

Reader	Precision	Target	Full	Adaptive-k	CAPE	K_{CAPE}
Qwen3-8B	BF16	Banking77	63.45/39.59	63.73/39.31	66.73/42.72	10.25
Qwen3-8B	4-bit	Banking77	63.58/39.27	63.40/37.90	66.72/41.92	10.59
Qwen3-8B	BF16	CLINC150	78.86/59.52	79.07/57.13	83.22/64.80	43.79
Qwen3-8B	4-bit	CLINC150	77.83/58.13	78.73/56.55	82.86/64.67	40.97

Table 20: **Deployment and offline cost audit.** Values cover all 72 local reader–target–suite settings. Input, output, and total tokens are query-weighted over all evaluated target queries and include target-schema costs amortized over the corresponding test set. Offline reader calls and tokens report reader-specific source fitting, macro-averaged over the 72 settings, and are excluded from target-time token counts. Compute-matched three-view costs are reported in Main Table 5.

Method	Calls/q	Schema calib. calls	Input tok/q	Output tok/q	Total tok/q	Offline reader calls	Offline reader tokens
Full	1	0	1,855	14	1,869	–	–
$K = 5$	1	0	317	9	326	–	–
$K = 10$	1	0	410	10	420	–	–
$K = 20$	1	0	569	11	579	–	–
Adaptive-k	1	0	474	10	484	–	–
LDAR	1	0	912	11	924	5,273	3.91M
CAPE	**3**	**87.11**	**1,982**	**34**	**2,017**	–	–

Relative to BF16, its Exact changes by -0.80 and -0.13 points, while its mean exposure changes by $+0.34$ and -2.82 intents. This control estimates within-reader precision sensitivity rather than a complete method ranking; the complete LDAR comparisons at each reader's deployed precision appear in Main Table 1 and Appendix Table 11.

K Cost and Amortization

For a deployment with Q real queries, let $B_{m,t}$ be the number of one-time target-schema calibration calls and n_{query} the number of reader calls for each real query. The amortized call count is

$$\text{Calls/query} = n_{\text{query}} + \frac{B_{m,t}}{Q}. \tag{61}$$

For native CAPE, $n_{\text{query}} = 3$: one canonical probe, one permuted probe, and one final prediction. For conditional arbitration in Appendix P, we replace n_{query} by the empirical mean of $C(x)$. Reported Calls/q counts real-query calls; one-time schema calls enter only the amortized total above. The naturalization generation cost is also one-time and must be reported separately because hosted APIs bill input and output tokens differently. LDAR performs no target-schema calibration, but it optimizes a source policy separately for each deployed reader and target. The offline columns report that reader-specific cost separately; it is excluded from both target calibration and target-time tokens.

The audit does not support describing CAPE as universally cheaper. CAPE pays three target-time calls and a one-time schema-calibration cost, whereas LDAR pays one target-time call but additionally averages 5,273 source-reader calls and 3.91M source-reader tokens for each reader–target–suite policy fit. Compute-matched three-view performance and token costs are reported in Main Table 5.

L Reasoning-Procedure Robustness

Table 21 separates the reasoning procedure from schema exposure in a controlled 2×2 comparison. Within each pair, the reader and reasoning instructions are unchanged; only Full versus CAPE exposure differs. The Direct comparison reports the effect of adaptive exposure under the common constrained classifier. The DSCP rows test whether the same exposure mechanism complements a distinct divide–solve–combine procedure.

Table 21: **Reasoning procedure × schema exposure.** F1 and Exact are macro-averaged over all 72 local reader–target–suite settings. Direct uses the common constrained classification prompt. Both DSCP rows retain its divide–solve–combine procedure; Full (native) is its original schema exposure, whereas the CAPE row changes only the displayed schema.

Reasoning procedure	Schema exposure	F1	Exact
Direct	Full	66.00	39.18
Direct	CAPE	70.72	44.91
DSCP	Full (native)	68.91	42.13
DSCP	CAPE	**73.44**	**47.85**

L.1 STRICT DSCP × EXPOSURE PROTOCOL

We fixed the following protocol before running the two DSCP conditions.

1. Generate the DSCP division once for each original utterance and cache it across the Full and CAPE exposure conditions.
2. Select one shared prefix for the original utterance; do not choose a separate depth for each divided clause.
3. Show every DSCP solve stage only the intents in the selected prefix $\mathcal{I}_K(x)$.
4. Restrict the combine stage to intent IDs in $\mathcal{I}_K(x)$; it cannot emit an intent outside the displayed prefix.
5. Share the coverage selector q_ϕ and immutable cosine ranking with the Direct condition.
6. Refit the reader-retention geometry on source-reader outcomes produced by DSCP rather than reuse geometry fitted from Direct outputs.
7. Run target-schema calibration through DSCP, using only synthetic description-derived probes and no target utterances or labels.
8. Compute canonical/permuted query stability from DSCP's final combined sets under identical pool membership.
9. Produce the final prediction by running DSCP with $\mathcal{I}_{K^*(x,m,t)}(x)$ after reader and query adaptation.

The DSCP + CAPE condition therefore uses source response geometry and test-time evidence obtained with DSCP, without reusing a depth selected under Direct. Adaptive exposure improves F1/Exact from 66.00/39.18 to 70.72/44.91 under Direct, corresponding to gains of 4.72/5.73 points. Under DSCP, performance increases from 68.91/42.13 to 73.44/47.85, with gains of 4.53/5.72 points. The similar improvements across reasoning procedures support the complementary role of adaptive exposure and indicate that its benefits extend beyond the Direct prompt.

M PREDICTIVE VALIDITY OF PRE-QUERY READER CALIBRATION

This audit asks whether reader adaptation measures behavior that is visible on real target queries. It freezes the pre-query, label-free reader posterior before any query-level order-agreement update and is therefore distinct from the final CAPE performance evaluation. Its reader evidence consists only of synthetic self-test outcomes. Unlabeled target-query coverage posteriors guide the allocation of remaining self-test calls, but no target label or reader output on a real query enters this pre-query posterior. Within each target–suite test split, we draw 400 distinct record indices uniformly without replacement using Python's `random.Random(seed+1000)` (`4407` for the reported seed), then sort the selected indices. This rule and sample size were fixed before any reader outcome was inspected; it uses no label, model output, stratification, or outcome-based replacement. The resulting 3,200 query IDs are shared by every reader, depth, and diagnostic method. This subset is used only to test response prediction and is not claimed to reproduce the magnitude of full-test performance changes; all headline task results use the complete test splits.

M.1 Evaluation Grid and Retention Event

For each query, the source-trained selector produces the predicted anchor $K_0(x)$ and coverage CDF $C_\phi(K \mid x)$. We evaluate the reader at the anchor and at the following target-specific grid $\mathcal{K}_t$:

Target	Evaluation depths
SNIPS	$\{1, 3, 5, 7\}$
ATIS	$\{1, 3, 5, 10, 15, 18\}$
Banking77	$\{1, 3, 5, 10, 20, 40, 77\}$
CLINC150	$\{1, 3, 5, 10, 20, 40, 80, 147\}$

Duplicate depths are executed once. The predicted anchor is retained even when it is not a member of this grid; the action set is

$$\mathcal{A}_x = \{K_0(x)\} \cup \{K \in \mathcal{K}_t : K > K_0(x)\}. \tag{62}$$

The minimum gold-cover rank is never substituted for K_0. Gold labels are read only after all predictions have been fixed.

For strict expansions, the capacity-derived prediction is the conditional retention proxy induced by Eq. 21:

$$p^{\text{proxy}}_{x,m,K} := \bar{S}_{m,t}(K \mid x) = \frac{\sum_{r=1}^{K} q_\phi(r \mid x) S_{m,t}(K - r)}{C_\phi(K \mid x)}. \tag{63}$$

For compactness, we write $p_{x,m,K} = p^{\text{proxy}}_{x,m,K}$ in the metric definitions below. We evaluate this pre-query proxy against observed correctness at larger depths, using only queries answered exactly at the predicted anchor. Thus, curve diagnostics include observations with $E_{K_0} = 1$ and $K > K_0(x)$. The anchor itself is excluded because its retention is one under this conditioning.

The proxy and the diagnostic use different conditioning events. The proxy conditions its rank mixture on gold coverage at K, while the diagnostic conditions on an exact prediction at K_0; these events are not assumed equivalent. All observed curves use the latter condition to evaluate the frozen proxy. Correctness is measured at each upper endpoint, so an intermediate error does not invalidate a later exact prediction.

M.2 Diagnostic References

We compare four prespecified estimates.

- **CAPE capacity** integrates endpoint retention over the pre-query reader posterior, before observing the real query's order-agreement signal.
- **Empirical stable depth** uses only synthetic self-tests that are exact at their anchor. It converts each test's largest uninterrupted stable expansion into a survival tail and averages those tails. This is a pathwise empirical control rather than the endpoint model used by CAPE.
- **Reader-blind cross-fit** averages observed target-response curves from the other eight readers. Five-fold query cross-fitting ensures that a query is predicted only from donor queries in the other folds. This reference uses target labels and is therefore a diagnostic upper-information reference, not a label-free deployment baseline.
- **Constant 0.5** assigns probability 0.5 to every strict expansion and is included only for Brier and NLL calibration context.

For a target–suite–reader–depth cell g, let $\mathcal{Q}_g$ contain the queries that satisfy the anchor-exact conditioning rule. Within this cell, $E_{x,K}$ denotes reader m's exact-prediction indicator for query x at depth K, and $p_{x,m,K}$ is its predicted retention. Let Ret_g and $\widehat{\mathrm{Ret}}_g$ denote the observed and predicted

mean retention:

$$\mathrm{Ret}_g = \frac{1}{|\mathcal{Q}_g|} \sum_{x \in \mathcal{Q}_g} E_{x,K}, \qquad \widehat{\mathrm{Ret}}_g = \frac{1}{|\mathcal{Q}_g|} \sum_{x \in \mathcal{Q}_g} p_{x,m,K}, \tag{64}$$

$$\mathrm{MAE}_g = \left| \widehat{\mathrm{Ret}}_g - \mathrm{Ret}_g \right|, \qquad \mathrm{Brier}_g = \frac{1}{|\mathcal{Q}_g|} \sum_{x \in \mathcal{Q}_g} (p_{x,m,K} - E_{x,K})^2, \tag{65}$$

$$\mathrm{NLL}_g = -\frac{1}{|\mathcal{Q}_g|} \sum_{x \in \mathcal{Q}_g} \left[E_{x,K} \log p_{x,m,K} + (1 - E_{x,K}) \log(1 - p_{x,m,K}) \right]. \tag{66}$$

NLL uses natural logarithms and clips probabilities to $[10^{-6}, 1 - 10^{-6}]$. Cells with fewer than 30 conditioned queries are excluded. MAE, Brier, and NLL are macro-averaged uniformly over depth, suite, target, and reader, in that order. Brier and NLL use individual-query probabilities, not cell means.

Reader-rank correlation is computed separately within each target–suite–depth condition as Spearman's ρ between readers' predicted and observed mean retention. Conditions with fewer than four valid readers or a constant vector are undefined and excluded before uniform aggregation over depth, suite, and target.

M.3 Controlled Selection and Regret

The following audit uses the finer query-level action set in Eq. 62 and measures per-query policy regret as a decision-oriented complement to the response-curve metrics above.

To isolate reader adaptation, let $p^{(h)}_{x,m,K}$ denote method h's retention estimate for a strict expansion. The controlled audit defines

$$G_h(K \mid x, K_0) = \begin{cases} 1, & K = K_0(x), \\ p^{(h)}_{x,m,K}, & K > K_0(x). \end{cases} \tag{67}$$

Thus the diagnostic anchor action receives score $C_\phi(K_0 \mid x)$. Each diagnostic estimate then selects

$$K_h(x) = \min \arg\max_{K \in \mathcal{A}_x} C_\phi(K \mid x) G_h(K \mid x, K_0), \tag{68}$$

where the minimum resolves score ties. This policy does not use query-level adaptation. Equation 67 is a normalization used only for this controlled audit; it is not the deployed CAPE score. In deployment, the anchor is scored by the same full convolution as every other action:

$$\widetilde{\mathcal{U}}(K_0 \mid x, m, t) = \sum_{r=1}^{K_0} q_\phi(r \mid x) S_{m,t}(K_0 - r), \tag{69}$$

which is generally not equal to $C_\phi(K_0 \mid x)$; the final method further replaces $p_{m,t}(c)$ by its query-updated posterior. The controlled policy's ex-post grid regret and Exact are

$$\mathrm{Regret}(h) = 100\, \mathbb{E}_x \left[\max_{K \in \mathcal{A}_x} E_{x,K} - E_{x,K_h(x)} \right], \tag{70}$$

$$\mathrm{Exact}(h) = 100\, \mathbb{E}_x [E_{x,K_h(x)}]. \tag{71}$$

Unlike the conditional curve metrics, these quantities include every sampled query, including queries that are wrong at the anchor. The query-grid oracle Exact is 45.90 over all nine readers and 43.15 over the six held-out readers. These are diagnostic action-grid results and must not be compared directly with final CAPE Exact, which uses complete test sets and query adaptation.

M.4 Example-Curve Cohort and Uncertainty

Figure 4(b) uses CLINC150 and two held-out readers selected solely by their pre-query capacity estimates before target outcomes were inspected: Llama-3.2-3B is the lower-tolerance reader and Qwen3-32B the higher-tolerance reader. The fixed cohort satisfies $K_0 < 5$ and requires both readers

Table 22: **Absolute predictive validity and decision quality of pre-query, label-free reader calibration.** Cells report estimate [95% query-bootstrap CI]. Curve metrics condition on an exact predicted-anchor response and strict expansion; policy metrics use all sampled queries. Regret is in percentage points. The held-out scope excludes the three Qwen readers used to fit response geometry. Reader-blind is a target-supervised, cross-fitted diagnostic reference rather than a deployment baseline. Dashes mark undefined quantities.

Diagnostic	Curve MAE ↓	Brier ↓	NLL ↓	Rank ρ ↑	Regret (pp) ↓	Policy Exact ↑
Six readers held out from response-geometry fitting						
Constant 0.5	–	0.250 [0.250, 0.250]	0.693 [0.693, 0.693]	–	–	–
Reader-blind cross-fit	0.218 [0.201, 0.239]	**0.220 [0.208, 0.236]**	**0.642 [0.611, 0.675]**	–	3.10 [2.57, 3.73]	40.05 [39.41, 40.68]
Empirical stable depth	0.231 [0.215, 0.254]	0.286 [0.265, 0.301]	1.170 [1.042, 1.321]	0.620 [0.493, 0.742]	3.95 [3.31, 4.61]	39.20 [38.48, 39.87]
CAPE capacity	**0.188 [0.174, 0.207]**	0.237 [0.224, 0.248]	0.668 [0.631, 0.689]	**0.762 [0.675, 0.839]**	**1.95 [1.52, 2.41]**	**41.20 [40.67, 41.74]**
All nine local readers						
Constant 0.5	–	0.250 [0.250, 0.250]	0.693 [0.693, 0.693]	–	–	–
Reader-blind cross-fit	0.211 [0.198, 0.229]	**0.216 [0.206, 0.231]**	**0.631 [0.605, 0.663]**	–	2.95 [2.49, 3.48]	42.95 [42.41, 43.47]
Empirical stable depth	0.218 [0.204, 0.239]	0.279 [0.266, 0.297]	1.120 [1.018, 1.242]	0.678 [0.591, 0.776]	3.75 [3.29, 4.30]	42.15 [41.58, 42.61]
CAPE capacity	**0.185 [0.170, 0.200]**	0.234 [0.217, 0.246]	0.661 [0.623, 0.687]	**0.768 [0.690, 0.846]**	**1.88 [1.45, 2.31]**	**44.02 [43.59, 44.45]**

Table 23: **Paired improvements of CAPE capacity over diagnostic references.** Each cell reports CAPE minus the named reference with a 95% paired query-bootstrap interval. Negative Curve MAE and regret, and positive rank correlation, favor CAPE.

Comparison	All nine	Held-out six
Curve MAE vs. Empirical	-0.033 [-0.053, -0.012]	-0.043 [-0.068, -0.019]
Curve MAE vs. Reader-blind	-0.026 [-0.046, -0.010]	-0.030 [-0.054, -0.007]
Reader-rank ρ vs. Empirical	$+0.090$ [$+0.042$, $+0.151$]	$+0.142$ [$+0.071$, $+0.224$]
Grid regret vs. Reader-blind (pp)	-1.07 [-1.49, -0.66]	-1.15 [-1.68, -0.61]
Grid regret vs. Empirical (pp)	-1.87 [-2.34, -1.39]	-2.00 [-2.64, -1.37]

to be exact at the anchor. It contains 64 Blend and 72 Mix queries, and the identical cohort is used at the five displayed depths $K \in \{5, 20, 40, 80, 147\}$. The shared-cohort and frozen-reader-selection checks both pass.

Confidence intervals use 10,000 paired query bootstraps (Efron, 1992). Within each target–suite fold, resampling a query carries all of its reader, depth, and method records together; reader-blind donor curves are recomputed in each replicate. The reported intervals condition on fixed readers, trained components, and schema calibration, and therefore quantify query uncertainty rather than training-seed or recalibration uncertainty. Of the candidate curve cells, 350 are valid for all nine readers and 232 for the held-out six; 28 cells fail the 30-query minimum in the all-nine scope. Reader-rank aggregation uses 36 and 34 valid conditions, respectively; six candidate conditions are undefined in the all-nine scope. All 10,000 non-rank bootstrap replicates and 9,950 valid rank replicates are reported, with no reader-blind or empirical fallback.

M.5 Observed Recovery on Real-Query Trajectories

The monotone boundary used for synthetic calibration is an approximation, not a claim that a frozen reader's correctness is monotone at every exposure on a real query. We quantify observed recovery on the same diagnostic outputs. Let $K_1 < \cdots < K_J$ denote the evaluated depths strictly above the predicted anchor. A query–reader trajectory is recovery-eligible when the reader is exact at the anchor, $J \geq 2$, and an incorrect endpoint occurs before the final evaluated depth:

$$E_{K_0} = 1, \qquad J \geq 2, \qquad \exists i < J : \; E_{K_i} = 0. \tag{72}$$

The last condition ensures that at least one higher evaluated depth remains at which recovery could be observed. An eligible trajectory is a recovery when there exist $i < j$ such that $E_{K_i} = 0$ and $E_{K_j} = 1$. Always-correct trajectories and trajectories whose first error occurs only at the final evaluated depth are excluded from the denominator. Each trajectory is counted once regardless of how many recovery pairs it contains.

For held-out readers, recovery rates range from 4.62% to 5.81% across targets, with a pooled rate of 5.47%. Recovery is therefore uncommon on the evaluated grids, even among trajectories with an error

Table 24: **Observed endpoint recovery on the prespecified diagnostic grids.** The denominator contains anchor-exact trajectories with at least two post-anchor depths and an error before the final evaluated depth, so that a subsequent recovery opportunity exists. A recovery is such a trajectory with an exact endpoint at a later depth. Rates are pooled over Blend/Mix and readers within each target; Macro averages the four target rates uniformly, whereas Pooled aggregates their counts. Intervals are 95% query-bootstrap CIs that retain all reader and depth observations for each sampled query. Held-out six is the primary scope.

	Six held-out readers		All nine local readers	
Target	Recovered/eligible	Rate (%) [95% CI]	Recovered/eligible	Rate (%) [95% CI]
SNIPS	3/65	4.62 [0.00, 10.77]	4/95	4.21 [0.00, 9.47]
ATIS	9/155	5.81 [2.58, 9.68]	14/235	5.96 [3.68, 8.81]
Banking77	14/260	5.38 [3.67, 7.54]	22/400	5.50 [4.01, 7.36]
CLINC150	21/380	5.53 [4.02, 7.39]	34/590	5.76 [4.52, 7.23]
Target macro	–	5.33 [2.74, 6.91]	–	5.36 [3.18, 6.75]
Pooled	47/860	5.47 [4.38, 6.71]	74/1320	5.61 [4.73, 6.58]

followed by an opportunity to recover. These results are consistent with using a monotone boundary as a compact schema-level approximation, while supporting evaluation of individual endpoints.

This conditional rate does not measure non-monotonicity across all trajectories. The prespecified grids may also miss recoveries between evaluated depths. Furthermore, the diagnostic uses real target utterances, whereas calibration estimates boundaries from synthetic self-tests. The result therefore provides supporting evidence without directly validating every synthetic boundary.

Table 25: **LDAR source-training budget sensitivity.** Performance cells report sample-F1/Exact (%). The LDAR columns denote total source-training utterances per leave-one-domain-out run. CAPE retains its original full-source training. Each reader row averages eight target–suite settings; Overall averages all 16 settings equally. ΔExact is CAPE minus LDAR-768 in percentage points, with a 95% paired query-bootstrap CI.

Reader	LDAR 192 train	LDAR 768 train	CAPE	ΔExact 95% CI
Qwen3-8B	78.92/59.45	79.08/59.81	79.61/60.71	$+0.90\,[-0.72, +2.81]$
Llama-3.2-3B	58.80/21.63	60.26/23.35	64.85/32.51	$+9.16\,[+7.61, +10.98]$
Overall	68.86/40.54	69.67/41.58	72.23/46.61	$+5.03\,[+3.88, +6.42]$

M.6 Setting-Level Exact-Transition Decomposition

Table 26 expands the aggregate audit in Table 6 to every local reader–target–suite setting. It is diagnostic only: all transition events are computed after inference from predictions, selected pools, and gold sets.

The detailed rows explain why aggregate gains coexist with a concentrated failure mode. For Qwen3-4B, 8B, 14B, and 32B on ATIS, the Blend/Mix mean coverage losses are 11.02, 14.12, 11.24, and 12.39 points, whereas the corresponding covered-pool losses are only 2.04, 2.33, 1.68, and 1.35 points. The dominant limitation in these settings is therefore candidate omission.

N LDAR Source-Training Budget Sensitivity

We evaluate a larger LDAR source-training budget on Qwen3-8B and Llama-3.2-3B, both excluded from CAPE's source response-geometry fitting. The evaluation covers four targets and both Blend/Mix suites, giving 16 reader–target–suite settings. For each leave-one-domain-out run, we increase LDAR's training set from 64 to 256 utterances per source domain, or from 192 to 768 utterances across the three source domains. The development set remains fixed at 32 utterances

Table 26: **Setting-level Full-to-Cape Exact-transition decomposition.** Entries are percentages of all test queries in the corresponding setting. G denotes Full wrong $\rightarrow$ Cape exact; L_{cov} denotes Full exact $\rightarrow$ Cape wrong with at least one gold intent absent from the selected pool; and L_{covered} denotes Full exact $\rightarrow$ Cape wrong despite complete gold coverage. Consequently, $\Delta = G - L_{\mathrm{cov}} - L_{\mathrm{covered}}$. Gold labels are used only for this post-hoc audit. †Reader used to fit source response geometry; ‡4-bit loading. Displayed entries are rounded to two decimals; the aggregate in Table 6 is computed from unrounded per-setting rates.

		Blend				Mix			
Reader	Target	G	L_{cov}	L_{covered}	Δ	G	L_{cov}	L_{covered}	Δ
Source Qwen readers (used to fit response geometry)									
Qwen3-0.6B†	SNIPS	27.84	6.62	1.91	+19.31	32.40	4.96	1.08	+26.36
	ATIS	8.27	0.98	3.11	+4.18	14.80	2.50	1.40	+10.90
	Banking77	5.20	0.10	6.83	-1.73	8.52	0.56	4.40	+3.56
	CLINC150	14.44	0.40	3.09	+10.95	8.11	0.45	5.08	+2.58
Qwen3-4B†	SNIPS	3.94	3.10	2.07	-1.22	3.00	0.92	2.36	-0.28
	ATIS	6.04	7.64	3.29	-4.89	4.90	14.40	0.80	-10.30
	Banking77	7.03	2.77	2.92	+1.34	7.91	5.79	2.67	-0.56
	CLINC150	8.33	1.34	2.52	+4.47	8.64	3.22	2.39	+3.03
Qwen3-14B†,‡	SNIPS	3.25	1.68	1.84	-0.27	3.36	3.16	0.88	-0.68
	ATIS	8.00	8.98	1.96	-2.93	9.40	13.50	1.40	-5.50
	Banking77	5.74	4.21	2.97	-1.43	6.35	6.24	2.45	-2.34
	CLINC150	5.11	0.84	2.08	+2.18	6.52	3.11	2.01	+1.40
Held-out Qwen sizes (unseen during response-geometry fitting)									
Qwen3-1.7B	SNIPS	10.90	5.47	3.33	+2.10	11.84	5.12	4.32	+2.40
	ATIS	6.76	2.49	3.20	+1.07	8.60	7.50	4.90	-3.80
	Banking77	8.61	0.94	3.07	+4.60	8.69	1.67	3.29	+3.73
	CLINC150	11.96	1.38	3.49	+7.09	12.35	2.01	3.45	+6.90
Qwen3-8B	SNIPS	2.49	1.30	2.41	-1.22	2.28	1.12	1.88	-0.72
	ATIS	6.49	9.24	2.76	-5.51	7.50	19.00	1.90	-13.40
	Banking77	10.49	3.17	3.07	+4.26	10.63	6.63	2.00	+2.00
	CLINC150	9.47	0.81	2.35	+6.32	8.90	1.74	2.92	+4.24
Qwen3-32B‡	SNIPS	2.68	1.72	1.19	-0.23	2.64	2.40	0.48	-0.24
	ATIS	10.22	10.58	1.60	-1.96	10.80	14.20	1.10	-4.50
	Banking77	4.16	2.92	1.68	-0.45	6.74	6.29	3.01	-2.56
	CLINC150	6.72	1.68	2.99	+2.05	6.74	5.04	2.99	-1.29
Held-out Llama readers (unseen model family)									
Llama-3.2-1B	SNIPS	39.16	2.64	0.76	+35.76	41.60	2.40	0.72	+38.48
	ATIS	11.20	0.09	0.27	+10.84	14.40	0.00	0.10	+14.30
	Banking77	11.28	0.00	0.05	+11.23	15.65	0.06	0.22	+15.37
	CLINC150	18.74	0.47	2.22	+16.06	13.79	0.27	2.65	+10.88
Llama-3.2-3B	SNIPS	24.93	5.35	2.29	+17.28	17.68	6.20	4.08	+7.40
	ATIS	17.16	1.24	1.78	+14.13	18.50	3.80	2.00	+12.70
	Banking77	14.65	0.54	0.94	+13.16	19.38	1.50	1.67	+16.20
	CLINC150	22.07	0.74	3.16	+18.17	23.19	2.43	1.82	+18.95
Llama-3.1-8B	SNIPS	13.38	5.77	5.51	+2.10	15.24	12.68	4.80	-2.24
	ATIS	16.00	2.84	3.11	+10.04	18.00	5.60	3.60	+8.80
	Banking77	13.41	1.68	1.48	+10.24	21.21	3.06	1.00	+17.15
	CLINC150	18.88	2.05	3.22	+13.60	20.12	3.79	3.03	+13.30

per source domain, or 96 in total. We retain the optimizer, stopping rule, and checkpoint-selection procedure described in Appendix D.6. CAPE's trained models and predictions remain unchanged.

For uncertainty estimation, we resample query IDs within each target–suite split, carrying all corresponding reader and method outcomes together. Each of 10,000 bootstrap replicates recomputes the equally weighted setting-macro difference. We report percentile 95% CIs, conditional on the fixed trained models, readers, and calibration results.

Increasing the source-training budget improves LDAR's macro Exact from 40.54 to 41.58, reducing CAPE's advantage from 6.07 to 5.03 points. The remaining difference is concentrated in Llama-3.2-3B (+9.16 points), whereas the Qwen3-8B interval includes zero. Thus, the macro advantage persists over the tested budget increase, but reader-level benefits remain heterogeneous. This two-reader analysis does not establish equal-data efficiency or saturation of LDAR's learning curve.

O Reader-Wise Gains and ATIS Selection Diagnostics

We examine the original CAPE policy. The analyses use the same nine local readers, four targets, and two suites (Blend and Mix), giving 72 reader–target–suite settings. The ATIS scope contains 18 settings. We average query-level quantities within each setting, then weight settings equally; conditional retention uses only gold-covered queries within each setting. Target labels are used only for evaluation and post-hoc diagnosis.

O.1 Gains by Reader and Target

Table 27 disaggregates the matched-budget comparison with reader-calibrated coverage thresholding and CAPE basic. Both controls use three-view aggregation, whereas full CAPE uses two evidence calls and one final prediction. For every reader, the point estimates favor full CAPE over both controls on Banking77 and CLINC150. Thus, the large-schema gains also occur among the larger Qwen readers and are not confined to small readers. ATIS is heterogeneous: full CAPE trails thresholding for Qwen3-4B, 8B, 14B, and 32B, while its target-macro difference against basic is -0.50 points. These cell-level point estimates do not establish significance for each reader–target pair.

Table 27: Exact differences by reader and target (percentage points). Each cell reports full CAPE minus K_τ / full CAPE minus basic, averaged over Blend and Mix. Both controls use three-view aggregation, with the same 64-call synthetic classification budget as full CAPE. (H) marks the six readers held out from fitting the source response geometry. Aggregate differences are computed before rounding.

Reader	SNIPS	ATIS	Banking77	CLINC150
Qwen3-0.6B	$+0.43/+0.14$	$+1.32/-0.48$	$+1.53/+1.14$	$+2.31/+1.83$
Qwen3-1.7B (H)	$+0.28/+0.06$	$+0.11/-0.39$	$+1.01/+0.86$	$+1.78/+1.41$
Qwen3-4B	$+0.18/+0.01$	$-0.79/-0.72$	$+0.94/+0.87$	$+1.72/+1.31$
Qwen3-8B (H)	$+0.03/-0.04$	$-1.16/-0.98$	$+0.79/+0.73$	$+1.64/+1.24$
Qwen3-14B	$+0.09/-0.06$	$-0.99/-0.87$	$+0.92/+0.84$	$+1.79/+1.12$
Qwen3-32B (H)	$-0.11/-0.14$	$-1.27/-1.04$	$+0.68/+0.67$	$+1.56/+1.18$
Llama-3.2-1B (H)	$+0.59/+0.21$	$+1.48/+0.13$	$+1.63/+1.12$	$+2.58/+1.79$
Llama-3.2-3B (H)	$+0.42/+0.11$	$+1.03/+0.03$	$+1.41/+1.09$	$+2.29/+1.73$
Llama-3.1-8B (H)	$+0.34/+0.07$	$+0.81/-0.18$	$+1.17/+1.05$	$+2.42/+1.80$
Target macro	$+0.25/+0.04$	$+0.06/-0.50$	$+1.12/+0.93$	$+2.01/+1.49$

O.2 Coverage and Retention at the Selected Depth

For each query, let $K^\star(x)$ be the final depth selected by the original policy and let R_x be the minimum rank that covers its complete gold set. Table 28 compares $C_\phi(K^\star(x) \mid x)$ with $\mathbf{1}\{R_x \le K^\star(x)\}$ over all queries. On ATIS, mean predicted coverage exceeds observed gold coverage by 21.7 points, demonstrating coverage overprediction at the selected depths.

Table 28: ATIS coverage at the final selected depth of the original CAPE policy. Predicted and observed coverage are percentages; the signed gap is predicted minus observed coverage in percentage points. Each reader row averages the two suites, and the macro equally weights all 18 settings. Aggregates and gaps are computed before rounding.

Reader	Predicted coverage	Gold coverage	Gap (pp)
Qwen3-0.6B	87.2	59.4	+27.8
Qwen3-1.7B	89.1	64.5	+24.6
Qwen3-4B	91.5	77.0	+14.5
Qwen3-8B	92.1	76.1	+16.0
Qwen3-14B	91.7	76.0	+15.7
Qwen3-32B	92.4	74.9	+17.5
Llama-3.2-1B	81.5	49.0	+32.5
Llama-3.2-3B	86.8	62.5	+24.3
Llama-3.1-8B	88.1	65.4	+22.7
Macro	88.9	67.2	+21.7

Let $S_x(d) = \sum_{c \geq d} p_{m,t}(c \mid z_x)$ denote the survival function of the query-updated capacity posterior. The conditional retention surrogate at a depth with positive predicted coverage is

$$\bar{S}(K \mid x, z_x) = \frac{\sum_{r=1}^{K} q_\phi(r \mid x) S_x(K-r)}{C_\phi(K \mid x)} = \frac{\widetilde{U}(K \mid x)}{C_\phi(K \mid x)}. \tag{73}$$

Table 29 evaluates both this surrogate and observed Exact on the identical gold-covered subset $\{x : R_x \leq K^\star(x)\}$, at each query's unchanged final depth. It is a post-update audit of the selected policy, not a pre/post comparison across changing query cohorts. It also differs from the pre-query, anchor-conditioned curve audit in Appendix M.

Table 29: ATIS retention on gold-covered queries at the final selected depth $K^\star(x)$. Predicted retention and observed Exact are evaluated on the same covered queries within each setting. Values are percentages; the signed gap is in percentage points. Each reader row averages the two suites; aggregates and gaps are computed before rounding.

Reader	Predicted retention	Covered-query Exact	Gap (pp)
Qwen3-0.6B	70.2	38.0	+32.2
Qwen3-1.7B	72.1	44.2	+27.9
Qwen3-4B	86.8	74.5	+12.3
Qwen3-8B	84.4	71.1	+13.3
Qwen3-14B	85.3	72.5	+12.8
Qwen3-32B	84.1	70.0	+14.1
Llama-3.2-1B	68.5	27.0	+41.5
Llama-3.2-3B	71.8	42.0	+29.8
Llama-3.1-8B	80.6	42.1	+38.5
Macro	78.2	53.5	+24.7

Because this surrogate mixes the capacity survival function with predicted cover ranks, this discrepancy cannot be attributed to capacity transfer alone. We therefore also evaluate a gold-rank substitution while retaining the same capacity posterior.

O.3 Gold-Rank Substitution at Fixed Policy Outputs

We replace $q_\phi(r \mid x)$ by a point mass at the observed gold-cover rank R_x, holding $K^\star(x)$, the query evidence z_x, the updated capacity posterior, and the final reader prediction fixed. The resulting diagnostic score is

$$U_{\text{gold-rank}}(K \mid x) = \begin{cases} S_x(K - R_x), & R_x \leq K, \\ 0, & R_x > K. \end{cases} \tag{74}$$

Both scores are evaluated at the original $K^\star(x)$, without reselecting the depth. This substitution uses target gold labels and is an evaluation-only diagnostic, not a deployable variant of CAPE.

Table 30 scores the original utility and the gold-rank diagnostic against the same binary Exact outcome on all queries, including uncovered queries. Query MAE is the mean absolute error between a score and the binary Exact outcome; Brier is the mean squared error, and NLL is binary log loss. The reported MAE is therefore distinct from the retention-curve MAE in Appendix M. Metrics are computed within each setting and then averaged with equal setting weights. Mean scores are percentages; the three error metrics use scores on the $[0, 1]$ scale.

Table 30: Utility-score diagnostics at unchanged CAPE policy outputs. Gold-rank substitution changes only the score used in this audit; selected depths, final predictions, and Exact remain unchanged. Both rows in each scope use the same queries and outcomes. All nine readers comprises 72 settings; ATIS comprises 18. Lower error metrics are better. The gold-rank row has access to target gold labels.

A. All nine readers				
Score	Mean (%)	Query MAE	Brier	NLL
Original $\widetilde{U}$	74.12	0.318	0.285	0.884
Gold-rank diagnostic	51.38	0.214	0.158	0.584
B. ATIS				
Original $\widetilde{U}$	77.24	0.412	0.365	1.127
Gold-rank diagnostic	42.85	0.221	0.168	0.562

With the capacity posterior unchanged, substituting the gold rank reduces Brier from 0.285 to 0.158 overall and from 0.365 to 0.168 on ATIS. Together with the direct coverage audit, these results identify cover-rank estimation as an important contributor to optimistic utility scores. Gold-rank substitution removes both rank uncertainty and estimation error, however, so its benefit does not isolate miscalibration alone. The residual discrepancy also remains nonzero. These diagnostics do not measure an achievable accuracy gain or establish that the capacity posterior is a calibrated predictor of real-query correctness. They distinguish score optimism from the decision value demonstrated by the matched-policy comparisons.

P Recovering ATIS Losses with Full-Default Arbitration

The preceding diagnostics link ATIS losses to gold omission and coverage overprediction. Full exposes every schema intent, allowing predictions unavailable within an underexposed proposal pool. The prediction union contains the complete gold set whenever either branch is exact, while keeping the verifier's pool small. Requiring confirmation before overriding Full is intended to limit harmful switches. This supplements exposure selection without correcting its coverage estimates; native CAPE remains the main method.

Protocol. The first call predicts $\widehat{Y}_P$ from the reader-adapted pool at $K_{\text{reader}}(x, m, t)$ in Eq. 10, before the query-level posterior update. A second call predicts $\widehat{Y}_F$ from the Full schema. If the two intent sets agree, we return their shared prediction. Otherwise, one verifier call classifies the query using only the descriptions in $U = \widehat{Y}_P \cup \widehat{Y}_F$, in a deterministic permuted order. For these disagreement cases, the final prediction is

$$\widehat{Y}_{\text{arb}} = \begin{cases} \widehat{Y}_P, & \widehat{Y}_P \neq \varnothing \text{ and } \widehat{Y}_V = \widehat{Y}_P, \\ \widehat{Y}_F, & \text{otherwise.} \end{cases} \tag{75}$$

Pool-local output IDs are first mapped back to shared intent identities; equality compares intent sets, independently of their displayed order. The final answer is one of the two branch predictions. No correctness signal is available during inference. Real-query calls satisfy

$$C(x) = 2 + \mathbf{1}\{\widehat{Y}_P \neq \widehat{Y}_F\}, \qquad \mathbb{E}[C] = 2 + \Pr(\widehat{Y}_P \neq \widehat{Y}_F). \tag{76}$$

Controls and accounting. Fixed-K and K_τ reuse the calibrated policies in Appendix I.2, changing only the proposal policy while retaining the Full branch, verifier, and override rule. Fixed-K selects one depth per reader–schema pair from the effective $\{5, 10, 20, N_t\}$ candidates; K_τ adapts depth to query coverage. The three policies share rankings, prompts, decoding, and the synthetic bank, with at most 64 synthetic classification calls per reader–schema pair; naturalization is counted separately. Full+3V uses canonical Full and two deterministic permutations with Dice-medoid aggregation as in Appendix D.6. The alternatives allow at most three real-query calls; native CAPE uses two probes and a final call. Token accounting and paired query-bootstrap CIs follow Appendices K and I.2, respectively.

Panel B adds a coverage-floor control, retaining native CAPE's query update and restricting Eq. 13 to $K \geq K_0(x)$ and $C_\phi(K \mid x) \geq 0.90$. We chose 0.90 after inspecting the ATIS results; this target-informed sensitivity control is exploratory and is not used for the headline results. The constraint concerns predicted coverage and does not guarantee gold inclusion.

Results and scope. Table 31 shows that arbitration raises ATIS Exact from 37.46 to 39.51, exceeding Full+3V by 0.80 points. It reverses native CAPE's deficits against Full on Qwen3-4B/8B/14B/32B, finishing 1.15/3.14/5.51/7.82 points above Full. Full+3V also recovers these deficits; arbitration's mean advantage over that control across the four readers is 0.26 points. On Qwen3-4B, Full+3V reaches 68.10 versus 66.12 for arbitration. Across all settings, arbitration exceeds Full+3V, fixed-K arbitration, and K_τ arbitration by 2.82, 2.80, and 1.45 points, respectively, with positive paired CIs also on held-out readers.

The coverage-floor control raises Exact over native CAPE by 1.48 points on ATIS and 0.28 overall. Arbitration exceeds this control by 0.57 points on ATIS but trails it by 1.21 overall. Thus, the stronger ATIS recovery comes with losses elsewhere.

Arbitration trails native CAPE by 0.93 points overall and 2.14 on held-out readers. SNIPS drops from 62.80 to 57.78 (-5.02 points), the largest negative contribution to the overall change; ATIS gains partly offset it. ATIS also shows large losses on Llama-3.2-1B/3B. Both query adaptation and the final decision rule change, so the causes of these losses remain unresolved. Arbitration is a complementary recovery experiment; a label-free choice between the two policies is not established, and native CAPE remains the main policy.

Table 31: **Full-default arbitration for ATIS recovery.** Exact is in percent. F denotes Full and V the union verifier. Panel A averages Blend/Mix; Panel B averages readers and suites and adds the 0.90 coverage-floor control. Panel C reports setting-macro accuracy and pool sizes over all 72 settings and the 48 held-out-reader settings; Δ is CAPE+F+V minus the row method, in percentage points, with paired-bootstrap 95% CIs. Bold marks the highest Exact among the methods displayed in each row or panel. Proposal $\bar{K}$ is the first-call pool size (the full schema for Full references and the final pool for native CAPE), not total exposure across calls. Calls/q is the observed real-query mean, with the maximum in parentheses; Tokens/q is query-weighted and includes amortized schema costs.

A. ATIS by reader

Reader	Full	Native CAPE	Full+3V	Fixed-K+F+V	K_τ+F+V	CAPE+F+V
Qwen3-0.6B	15.03	**22.57**	17.61	15.92	17.06	21.68
Qwen3-1.7B	29.88	28.51	**35.19**	30.61	31.25	30.34
Qwen3-4B	64.97	57.37	**68.10**	65.18	66.31	66.12
Qwen3-8B	63.57	54.11	65.07	64.23	65.38	**66.71**
Qwen3-14B	59.32	55.10	64.60	61.87	62.94	**64.83**
Qwen3-32B	55.66	52.43	62.33	58.74	60.47	**63.48**
Llama-3.2-1B	0.66	**13.23**	0.09	0.83	0.87	1.15
Llama-3.2-3B	12.83	**26.25**	14.39	14.42	15.03	15.63
Llama-3.1-8B	18.11	**27.53**	21.01	21.56	21.07	25.65
ATIS macro	35.56	37.46	38.71	37.04	37.82	**39.51**

B. By target and overall

Target	Full	Native CAPE	CAPE +floor	Full +3V	Fixed-K +F+V	K_τ +F+V	CAPE +F+V
SNIPS	54.80	**62.80**	62.73	56.26	55.18	55.87	57.78
ATIS	35.56	37.46	38.94	38.71	37.04	37.82	**39.51**
Banking77	26.45	**31.66**	31.60	27.95	28.93	29.96	31.54
CLINC150	39.90	**47.73**	47.49	41.72	43.57	46.47	47.09
Macro	39.18	44.91	**45.19**	41.16	41.18	42.53	43.98

C. Aggregate accuracy, paired differences, and cost

Method	All-9 Exact	Held-out-6 Exact	All-9 ΔExact [95% CI]	Held-out-6 ΔExact [95% CI]	Proposal $\bar{K}$	Calls/q	Tokens/q
Full	39.18	34.34	$+4.80$ $[4.24, 5.37]$	$+5.25$ $[4.48, 6.03]$	62.25	1.00 (1)	1,869
Full+3V	41.16	36.11	$+2.82$ $[2.31, 3.37]$	$+3.48$ $[2.73, 4.21]$	62.25	3.00 (3)	5,607
Fixed-K+F+V	41.18	36.46	$+2.80$ $[2.18, 3.35]$	$+3.13$ $[2.37, 3.94]$	15.34	2.41 (3)	2,567
K_τ+F+V	42.53	37.75	$+1.45$ $[0.88, 1.97]$	$+1.84$ $[1.03, 2.61]$	12.80	2.47 (3)	2,518
Native CAPE	**44.91**	**41.73**	-0.93 $[-1.51, -0.34]$	-2.14 $[-3.05, -1.28]$	11.27	3.00 (3)	2,017
CAPE+F+V	43.98	39.59	Reference	Reference	8.71	2.59 (3)	2,482