

How Much Evidence Should Retrieval-Augmented In-Context Learning Use Under Distribution Shift?

Chen Wang^{1,2,*}

¹College of Systems Engineering, National University of Defense Technology, Changsha 410073, Hunan, China

²National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, Hunan, China (*Corresponding author)

TL;DR: Retrieval helps only while aligned evidence grows faster than reader-facing ambiguity, which explains recall–accuracy decoupling under shift.

Presentation Info

Sydney Poster Session 1

Venue: Sydney, Australia
Exhibition Hall 1–4

Date: Tue, Dec 8, 2026

Time: 10:00–13:00 AEDT
(07:00–10:00 Beijing)

1. Problem & Core Law

The Budget Dilemma: Under distribution shift, increasing top- k improves **evidence coverage** ($\eta_{\text{ret}}r$), but also injects **distractors, label conflict, and length overhead** (ε_k).

Effective Evidence Coverage Law:

$$\text{Risk}_n^{\text{RAG}} \approx \text{base} + c_3 \left[(\gamma - \eta_{\text{ret}}r)_+ + \varepsilon_k \right] \sqrt{\alpha\beta}$$

Principle: Retrieval reduces risk only while aligned evidence grows faster than reader-facing ambiguity.

2. Falsifiable Signature & Regimes

Recall $\neq$ Accuracy: Gold-evidence recall can keep rising after answer accuracy saturates or falls.

Three Diagnostic Budget Regimes:

- Coverage** SQuAD, MS MARCO: rises with k ; rerank/compress.
- Saturation** SciQ, PubMedQA: top-1 suffices; extra k wastes tokens.
- Interference** FEVER, MNLI, HotpotQA: extra k hurts; cap/filter.

Cheap Diagnosis: The sign of Fixed-4 – Fixed-1 separates the three regimes across readers without training.

3. Validation Scope & Probes

Two-Tier Validation:

- Linear-ICL Synthetic:** Exact check of linear- γ coupling ($R^2=0.999$) and interior k^* (15/15 cells).
- Frozen LLMs (0.5B–7B + API):** Qwen2.5, DeepSeek-V3.1, Kimi-K2.5, Gemini-3.6-Flash, Qwen3.7-Max.

Operational Probes (Zero Training): RA-EEB, QCA-EEB, and Router-EEB test whether the diagnosis is actionable using frozen calibrated thresholds.

Disclosure: Sole author; no third-party funding; no competing financial or non-financial interests.

Motivation & Effective Evidence Coverage Paradigm

Why retrieval-augmented in-context learning under distribution shift is a budget problem, not a recall-only problem

NeurIPS 2026 · #36573 | Slide 2 / 6

Core Mechanism & Diagnostic Commitments

Why Single-Shift Theory Is Insufficient:

- Covariate shift (Δ_x) and task-prior shift (Δ_w) interact through a resolvent-weighted mixed term $c_3\gamma\sqrt{\alpha\beta}$.
- Single-shift analyses are additive in α and β , so they cannot explain why two tasks under the *same* retriever and reader follow opposite budget trends.

Competition Between Coverage and Ambiguity:

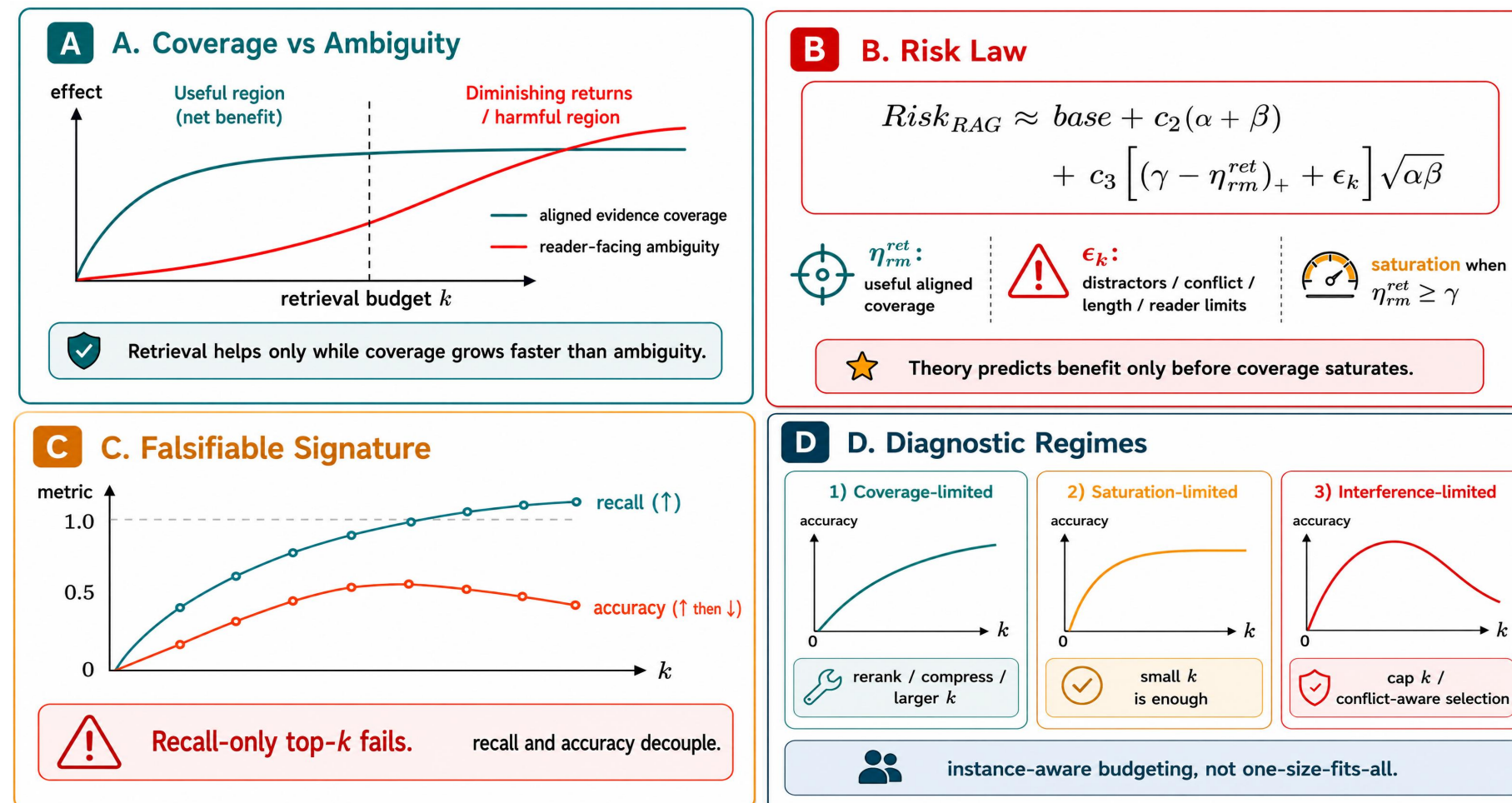
- Aligned Coverage (η_{ret}^r):** Retrieval along the harmful coupled-shift direction cancels cross-shift risk until saturation ($\eta_{\text{ret}}^r \geq \gamma$).
- Reader-Facing Ambiguity (ϵ_k):** Off-direction distractors (ϵ_{dist}), prompt length (ϵ_{len}), and finite reader capacity (ϵ_{read}) accumulate with k .

Explicit Falsification Commitments:

- P1 (Decoupling):** Falsified if accuracy is monotone in recall after controlling prompt length.
- P2 (Semantics vs. Length):** Falsified if oracle fails to recover accuracy or shuffled same-length context matches retrieved context.
- P3 (Regime Stability):** Falsified if regime labels invert erratically across nearby reader scales.

Scope: Spectral quantities ($\gamma, \rho, \eta_{\text{ret}}$) are provable in local linear ICL and treated as *latent* in frozen LLMs.

Overview of the Theory-Guided Budget Diagnosis (Figure 1)



Reading Figure 1:

- (A) Marginal Tradeoff:** Retrieval helps only while marginal aligned coverage exceeds marginal ambiguity.
- (B) Coupled-Shift Law:** Separates useful shift-aligned coverage η_{ret}^r from ambiguity overhead ϵ_k .
- (C–D) Signature & Regimes:** Gold recall rises after accuracy peaks, splitting tasks into Coverage, Saturation, and Interference regimes.

Local Linear-ICL Theory & Synthetic Validation

Direct quantitative verification where spectral coordinates $(\alpha, \beta, \gamma, \rho, \eta_{\text{ret}})$ are controlled

NeurIPS 2026 · #36573 | Slide 3 / 6

Theorems 1–3 & Quantitative Synthetic Checks

Thm. 1–2 (Coupled Shift & Breakdown) & Thm. 3 (Coverage Law):

$$\text{Risk}_n^{\text{RAG}}(\alpha, \beta, \gamma; r) = \frac{c_1 d}{n} + c_2(\alpha + \beta) + c_3[(\gamma - \eta_{\text{ret}} r)_+ + \varepsilon_k] \sqrt{\alpha\beta} + O((\alpha + \beta)^{3/2} + r^2)$$

where $\alpha = \|\Delta_x\|_{\mathcal{M}}^2$, $\beta = \|\Delta_w\|_{\mathcal{M}}^2$, $\gamma \in [-1, 1]$ is shift alignment, $r = k\rho^{-2}$ is coverage, and $\eta_{\text{ret}} \in [0, 1]$ is aligned projection.

Synthetic Experiment	Lin.- γ R^2	Even- γ R^2	Cells	Key Finding
1a: PSD shift ($d=n=60$)	0.621 (0.84 [†])	0.288	9/9	Monotone in γ ; $\gamma < 0$ cancels up to 96% excess risk
1b: Trace-free ($d=n=60$)	0.999	0.990	9/9	Zero 1st-order term; confirms linear $\gamma\sqrt{\alpha\beta}$
2: Rank-decay ($d=20, n=48$)	$p_j = p_0 e^{-(j-1)/8}$		15/15	Interior $k^* \in [4, 10]$; 6.9%–14.2% post- k^* rise
2b: Conflict sweep ($c=0.4/1.0$)	boundary shifts left		14/15, 8/15	Rise 10.7%–20.9% at $c=0.4$; else $k^*=0$

Mechanism Division: Optimum k^* tracks retriever precision ($k^*=6, 7, 9$ for $p_0=0.5, 0.7, 0.9$ at $\gamma=0$), while coupling γ sets the post-boundary penalty ($0.100 \rightarrow 0.236$ risk units). [†]With 1st-order terms.

Interior Budget Boundary & Recall–Risk Decoupling

15 / 15 Cells

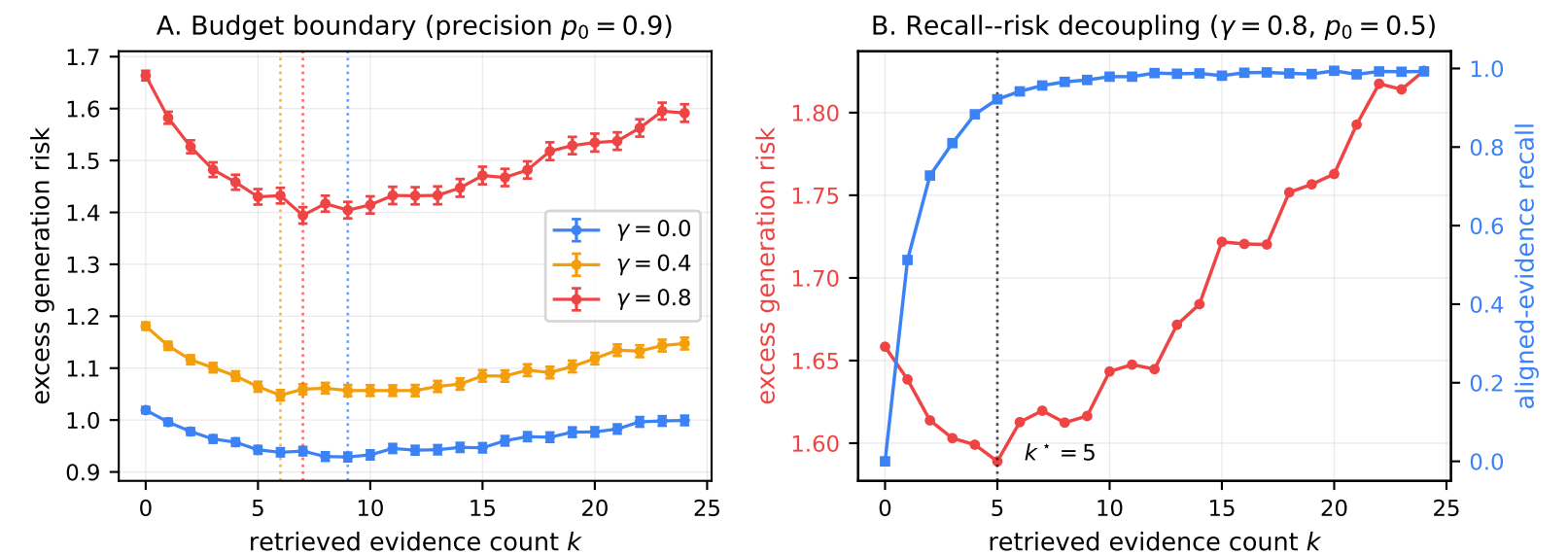
Interior optimum $k^* \in [4, 10]$
under rank-decaying p_j

+6.9% to +14.2%

Post- k^* excess risk rise
as coupling γ increases

+10.0% to +14.9%

Risk penalty when pushing
recall from 0.92 to 0.99



Key Synthetic Findings:

- Left (Budget Boundary):** Under rank-decaying precision $p_j = p_0 e^{-(j-1)/8}$ ($\tau=8$), an interior optimum $k^* \in [4, 10]$ appears in 15/15 cells, and post- k^* excess risk rises monotonically with coupling γ .
- Right (Recall–Risk Decoupling):** At $p_0=0.5$, the risk optimum sits at aligned recall 0.89–0.95; pushing retrieval to full recall (0.99) increases risk by **10.0%–14.9%**.

Conditional Reader Transfer & Real-System Budget Sweeps

Testing Proposition 4, Recall–Accuracy decoupling (P1), and Oracle vs. Shuffled controls (P2)

NeurIPS 2026 · #36573 | Slide 4 / 6

Proposition 4 & Theory-to-System Control Map

Proposition 4 (Conditional Transfer to Frozen Readers): Under prompt-level stability $\mathcal{L}_M(k) = L_{0,M} + a_M(\gamma - c_k)_+ + b_MA_k + \zeta_k$ ($|\zeta_k| \leq \Delta_M$), increasing budget from k_1 to $k_2 > k_1$ hurts performance whenever ambiguity grows faster than coverage:

$$b_M(A_{k_2} - A_{k_1}) > a_M((\gamma - c_{k_1})_+ - (\gamma - c_{k_2})_+) + 2\Delta_M \implies \mathcal{L}_M(k_2) > \mathcal{L}_M(k_1).$$

Theory Term	System Proxy	Control	Falsified If → Observed Outcome
η_{rer}^r (coverage)	Gold recall / rerank	Oracle context	Oracle fails → confirmed: 0.700 → 0.863 (+16.3 pt, Gemini FEVER)
$\varepsilon_{\text{dist}}$ (conflict)	Stance entropy H_k	NLI verifier, dense- k	Monotone gain → confirmed: FEVER/MNLI best at (anchored) top-1
ε_{len} (length)	Prompt token count	Same-length shuffle	Shuffle matches → confirmed: 0.46–0.60 (−23.7 pt vs. no context)
$\varepsilon_{\text{read}}$ (capacity)	Reader scale (0.5B–7B)	Scale & API sweep	Random flip → confirmed: QA 0/10 vs. NLI/Verif. 6/12 interference

P2 Dissociation Confirmed: On Gemini-3.6-Flash FEVER ($n=300$), oracle passage insertion lifts accuracy +16.3 pt over no context and +8.6 pt over best retrieved k , whereas length-matched shuffled context collapses accuracy by up to −23.7 pt.

Headline Dense- k Sweeps & Token Tradeoffs (Figure 2)

Coverage (SQuAD)

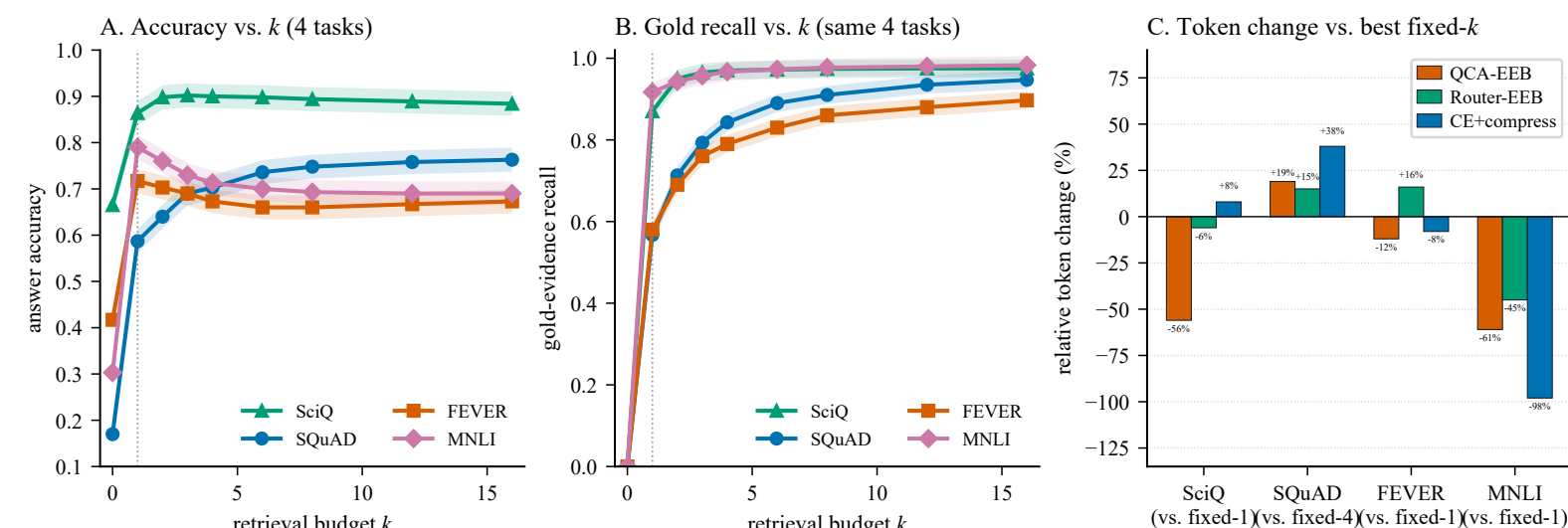
Accuracy climbs to $k=16$
(0.271 → 0.809 on DeepSeek)

Saturation (SciQ)

Plateaus at top-1
(F4 − F1 = +0.001)

Interference (FEVER)

Peaks at $k=1$, degrades for $k \geq 4$ (F4 − F1 = −0.048)



P1 Confirmed Across Headline Tasks:

- **(A–B)** Gold recall rises monotonically with k on all four tasks, but accuracy splits into three regimes: **SQuAD** climbs (coverage-limited), **SciQ** plateaus early (saturation-limited), and **FEVER/MNLI** peak at small k then degrade (interference-limited).
- **(C)** Selected-context probes adjust prompt tokens according to regime rather than applying a one-size-fits-all budget.

Cross-Scale Regime Matrix & Large-Sample Robustness

Regime stability across open-weight scales (0.5B–7B), frontier API readers, and $n = 1000$ reruns

NeurIPS 2026 · #36573 | Slide 5 / 6

Regime Stability Across Scales & Families (Figure 3)

0 / 10 QA Cells

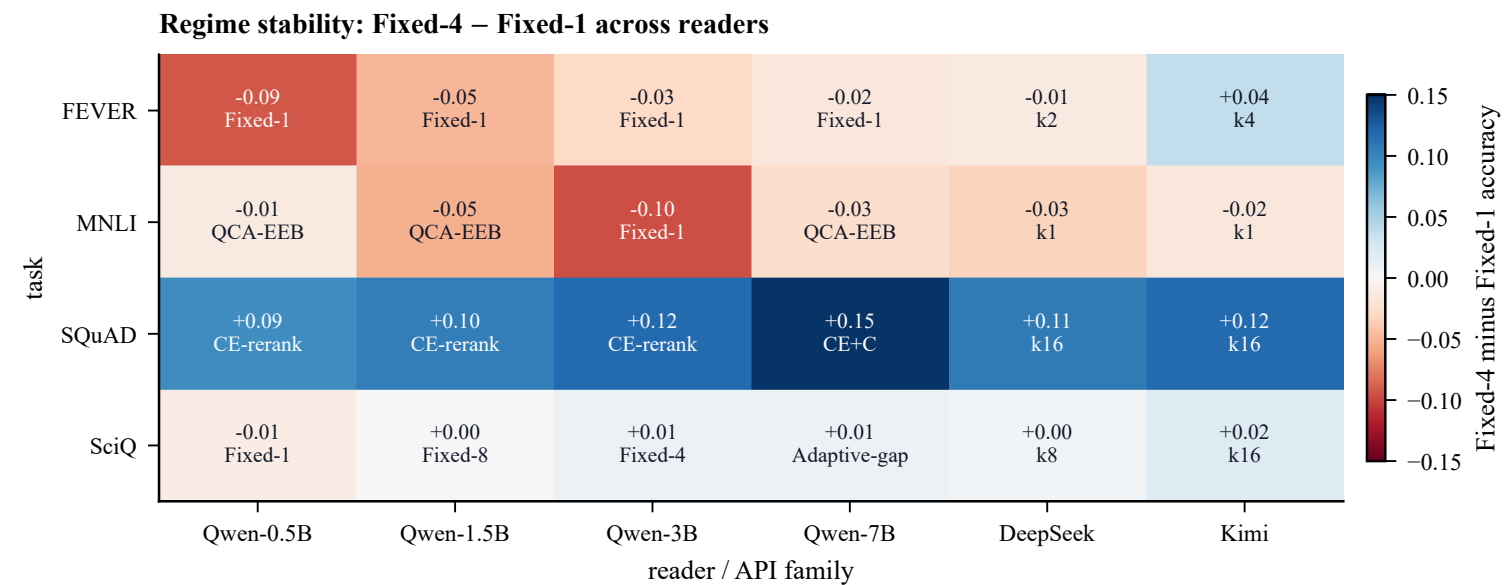
Interference-limited
above 0.5B scale

6 / 12 NLI/Verif.

Interference-limited
across scales & APIs

$1.76 \times -6.1 \times$

Extra tokens spent at
 $k=16$ for ≤ 0 gain



Asymmetric Task-Family Split:

- Above 0.5B, QA is interference-limited in **0/10 cells** (DeepSeek-V3.1 SQuAD at $n=1000$ rises monotonically 0.271 $\rightarrow$ 0.809).
- Verification/NLI is interference-limited in **6/12 cells**, and **12/14 cells** reach optimal accuracy at $k^* \leq 8$: retrieving to $k=16$ spends **$1.76 \times -6.1 \times$ tokens** for ≤ 0 gain.

Reviewer-Strengthening Diagnostics (Figure 5)

$n = 1,000$ Reruns

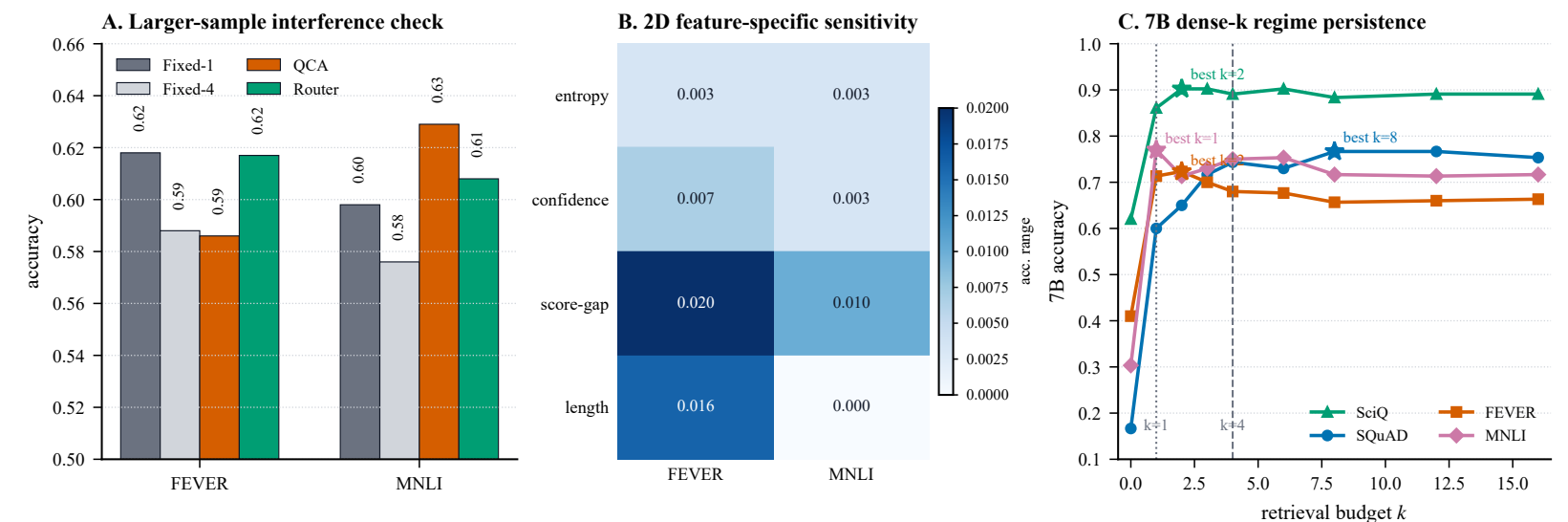
FEVER & MNLI interference
persists at large sample size

≤ 0.020 Range

2D threshold sensitivity on
FEVER (≤ 0.010 on MNLI)

7B Dense- k Check

SQuAD cover-
age vs. FEVER
interference holds at 7B



Large-Sample & Sensitivity Verification:

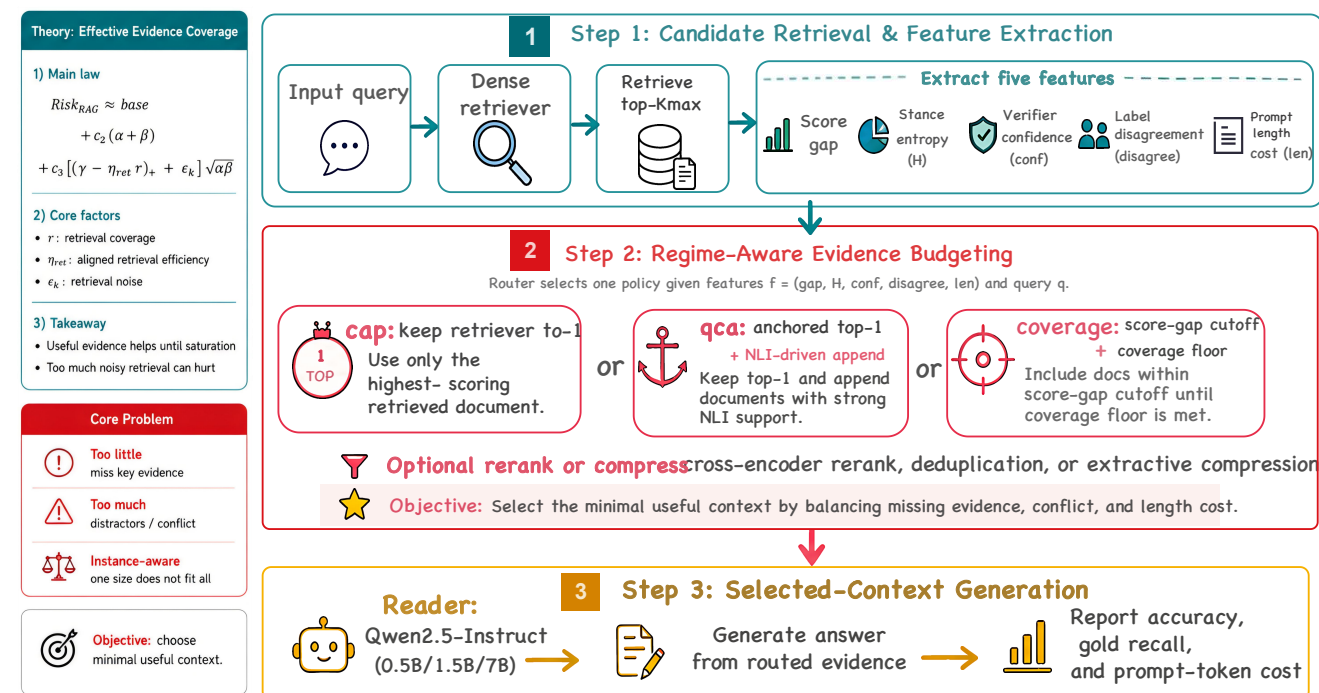
- (A) $n=1000$ Rerun:** Interference pattern on FEVER and MNLI persists at $n=1000$ beyond the $n=300$ headline runs.
- (B) 2D Threshold Perturbation:** Accuracy range is ≤ 0.020 on FEVER and ≤ 0.010 on MNLI under single-feature perturbations.
- (C) 7B Dense- k Curves:** SQuAD stays coverage-limited while FEVER remains interference-limited at 7B.

Operational Controller Probes, Stability Proxy & Takeaways

Zero-training selected-context probes (RA-/QCA-/Router-EEB) and Proposition 4 empirical proxy check

NeurIPS 2026 · #36573 | Slide 6 / 6

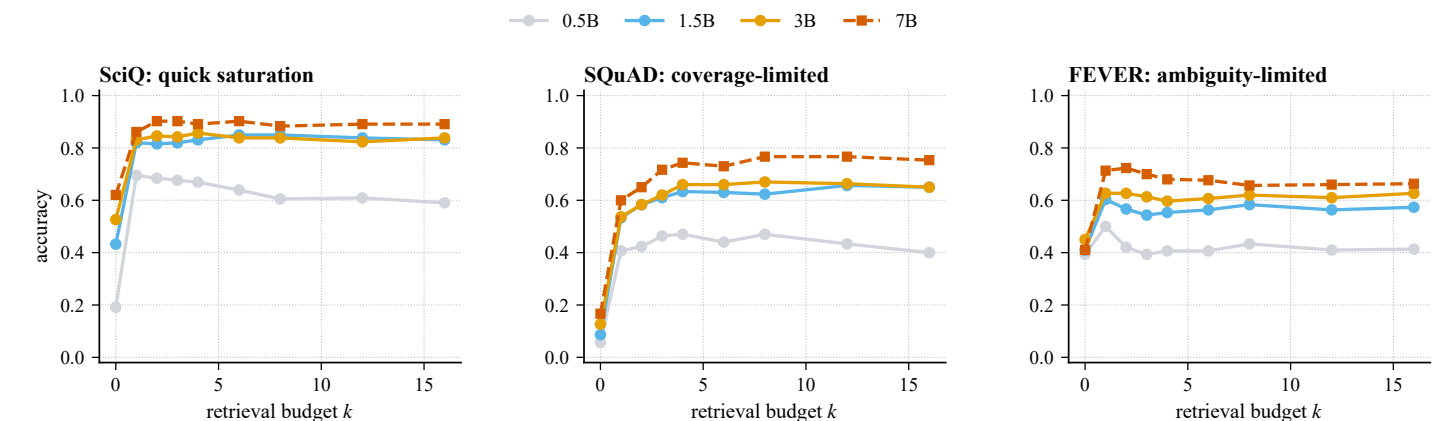
Selected-Context Probe Pipeline & Regime Policy Map



Task	Regime	F4–F1	Actionable Policy	Best Policy (0.5B/1.5B/3B/7B)
SQuAD v1.1	Coverage	+0.115	Rerank / compress	CE rerank (CE+C at 7B)
MS MARCO	Coverage	+0.032	Larger aligned k	Fixed-4 or adaptive gap
SciQ	Saturation	+0.001	Small fixed k	Near-flat; winner varies
FEVER	Interference	−0.048	Cap k / conflict filter	Fixed-1 at all 4 scales
MNLI	Interference	−0.048	Cap k / stance filter	QCA (3/4) or Fixed-1
HotpotQA	Interference	−0.057	Cap k	Fixed-1 at all 4 scales

Role of Probes: Thresholds $\Theta = (\tau_g, \tau_H, \tau_c, \tau_d, \tau_\ell)$ are calibrated once on a disjoint split and frozen. Router-EEB ties the strongest baseline on 10/12 headline cells.

Multi-Scale Curves, Prop. 4 Proxy & Takeaways



Frozen Reader	a (miss)	b (entropy)	$\hat{\Delta}_M$ [95% CI]	Ratio	Diagnostic Reading
Qwen2.5-0.5B	0.127	0.071	0.034 [0.019, 0.058]	0.314	Loosest proxy fit
Qwen2.5-1.5B	0.287	0.071	0.025 [0.014, 0.050]	0.129	Strong coverage slope
Qwen2.5-3B	0.283	0.061	0.024 [0.010, 0.043]	0.138	Ambiguity slope eases
Gemini-3.6-Flash	0.093	0.017	0.015 [0.006, 0.035]	0.209	Mild ambiguity slope
Qwen3.7-Max	0.079	−0.015	0.014 [0.006, 0.033]	0.136	Absorbs conflict

Sufficient harmful-increase condition on FEVER ($n=300$): fired / violated = 0/0 for all readers.

Key Takeaways:

- Recall $\neq$ Accuracy under shift:** Higher recall can reduce answer accuracy once coverage saturates.
- Regime-guided policy:** Rerank/compress for coverage QA; cap k or filter conflict for verification/NLI.
- Disclosure:** Sole author; no third-party funding; no competing interests.