Where Should Evidence Intervene? Proposal-Admission Decoding for Grounded Answer Sets

Anonymous ACL submission

Abstract

In grounded answer-set generation, an F1 gain is interpretable as grounded improvement only when it depends on the matched retrieved evidence. This requirement is especially important for masked diffusion language models, where evidence contrast perturbs coupled denoising trajectories and can jointly affect which answer slots emerge and which enter the final answer set. We propose Proposal-Admission Decoding, a two-stage decoding interface that separates contrastive candidate generation from slot-level evidence-local admission for final answer-set inclusion. On ALCE-QAMPARI with LLaDA-8B-Instruct, trajectory-level contrast is useful for surfacing candidates but unstable as a final answer-set boundary. Holistic factuality scorers obtain strong F1 gains, yet these gains persist under shuffled evidence. In contrast, Proposal-Admission Decoding with MiniCheck-Flan and WeCheck yields positive F1 gains that collapse under shuffled evidence, indicating matched-evidence dependence. We summarize this distinction with a binding-ratio diagnostic. These results support a boundary principle for grounded answer-set decoding: contrastive trajectories can expand the candidate set, while final answer-set inclusion should be licensed at a slot-level evidence-local boundary.

1 Introduction

Grounded answer-set generation requires models to produce a set of answers that are supported by retrieved evidence. In benchmarks such as QAMPARI (Gao et al., 2023; Amouyal et al., 2022), the system receives a query and retrieved passages, and must output multiple valid answer entities rather than a single span. Because the output is a set, grounded decoding involves two distinct stages: the model must surface plausible candidate answers, and it must decide which candidates are sufficiently supported to enter the final answer set. We refer to these stages as *candidate generation* and *final inclusion*. Distinguishing them is important because candidate discovery and evidence-grounded acceptance can otherwise be conflated.

This distinction becomes especially consequential for masked diffusion language models (MDLMs). Unlike autoregressive decoders, which generate one next token at a time, MDLMs repeatedly refine many masked positions across denoising steps (Li et al., 2022; Sahoo et al., 2024; Nie et al., 2025). Retrieved evidence can therefore influence multiple positions and multiple intermediate denoising states before the final output is realized. A single evidence-contrast signal may affect both which answer candidates emerge and which remain in the final answer set.

This creates an evidence-intervention boundary problem for retrieval-augmented diffusion decoding. The central question is not only whether retrieved evidence improves answer-set F1, but at which stage evidence should affect the decoding decision. For answer-set generation, this placement matters because a method may use evidence to surface better candidates while still failing to decide which candidates should enter the final answer set. Generation-time contrastive decoding (Shi et al., 2024; Kim et al., 2024; Wang et al., 2025; Yu et al., 2026) injects contextual or evidential signals into per-step logits or denoising updates. In doing so, the same decoding process is responsible both for proposing candidate answers and for deciding final answer-set membership. As a result, an improvement in the final answer-set F1 may reflect a more plausible or higher-recall output, without necessarily depending on the matched retrieved evidence.

Sequence-level factuality filters (Zha et al., 2023; Laban et al., 2022) do not fully resolve this problem either. They assign a global plausibility score to each complete output, but do not directly test whether each candidate is locally supported by evidence for the query. This makes them useful as

global plausibility filters, but not as a direct test of slot-level evidence binding. Indeed, we show that such filters can retain their apparent gains even when evidence is shuffled across examples, indicating that their improvements are not tightly bound to the matched retrieved evidence.

To address this problem, we propose Proposal-Admission Decoding, a two-stage decoding interface that separates contrastive candidate generation from evidence-bound final inclusion. A permissive contrastive proposer surfaces candidate slots from a contrast-shaped trajectory, while a slot-level admission rule decides which slots are supported by matched evidence and can enter the final answer set. We instantiate this interface with SPREAD-style relevance (Yu et al., 2026) as the proposer and small NLI verifiers as admission models.

On QAMPARI, we use Proposal-Admission Decoding in a conservative repair setting: the contrastive proposer is invoked only when the retrieved-evidence baseline returns no answer slots, while populated baselines are left unchanged. This setting isolates the central boundary question: whether contrastive diffusion can recover evidence-supported candidates without letting the denoising trajectory decide final answer-set inclusion. A non-empty ablation further tests the same proposal-admission boundary under broader routing.

On ALCE-QAMPARI with LLaDA-8B-Instruct, this boundary view changes the interpretation. In particular, official F1 alone conflates candidate discovery, global plausibility filtering, and matched-evidence-bound final inclusion. Proposal-only decoding improves official F1 by +0.539, showing that contrastive diffusion can supply missing candidates, but every proposed slot is accepted by construction. Holistic scorers improve official F1, but their gains persist under shuffled evidence; SummaCConv retains its full +0.542 gain.

By contrast, Proposal-Admission Decoding improves official F1 by up to +0.360, while its shuffled-evidence gain collapses to 0.000. Thus, although slot-level admission yields smaller gains than the strongest holistic scorer under official F1, its gains collapse under shuffled evidence, indicating dependence on matched retrieved evidence.

In summary, we make three contributions.

- **Boundary diagnosis for diffusion-based grounded decoding.** We identify an evidence-intervention boundary problem in retrieval-augmented masked diffusion decoding: trajectory-level evidence contrast can improve candidate discovery while leaving final answer-set inclusion insufficiently tied to matched retrieved evidence.

- **Proposal-Admission Decoding.** We instantiate a two-stage decoding interface for grounded answer sets. A contrastive diffusion proposer expands the candidate slot set, while slot-level evidence-local admission decides which candidates are included in the final answer set.

- **Matched-evidence dependence evaluation.** We introduce fixed-candidate shuffled-evidence controls and a binding-ratio diagnostic to distinguish evidence-bound improvements from evidence-leaky F1 gains. This reveals that holistic scorers can improve official F1 while retaining gains under mismatched evidence, whereas slot-level admission produces gains tied to matched evidence.

2 Related Work

2.1 Masked Diffusion Language Models

Masked diffusion language models (MDLMs) generate text by iteratively refining masked positions, rather than extending a left-to-right prefix (Li et al., 2022; Sahoo et al., 2024; Nie et al., 2025). Because many positions are updated across multiple denoising steps, retrieved evidence can affect many parts of the output before the final sequence is produced. This makes the placement of evidence especially important: an evidence signal injected during denoising may change both which answer slots appear and which slots remain in the final output. Recent retrieval-augmented diffusion decoders often intervene directly in the denoising process. For example, SPREAD (Yu et al., 2026) uses query relevance to reduce semantic drift in diffusion-based RAG. Our work uses this kind of relevance-guided decoding to surface candidate slots, but leaves final answer-set inclusion to a separate admission step.

2.2 Contrastive Decoding

Context-aware decoding (Shi et al., 2024) amplifies the difference between with-context and without-context distributions during token sampling, while AdaCAD (Wang et al., 2025) and Adaptive Contrastive Decoding (Kim et al., 2024) adaptively

scale contextual influence. PCED (Corallo and Papotti, 2026) treats retrieved documents as separate experts and moves multi-document evidence aggregation to a retrieval-aware contrastive decoding rule. More broadly, DoLa (Chuang et al., 2024) contrasts the output distributions of different transformer layers. In the diffusion-LM setting, classifier-free guidance variants such as A-CFG (Li et al., 2025) and SNR-guided ARAM (Kim and Ye, 2026) apply guidance across denoising steps. These methods differ in what they contrast and how strongly they apply the contrast, but they use contrast during generation and take the resulting sequence as the final output. In contrast, Proposal-Admission Decoding uses contrastive decoding only to propose candidate slots; whether those slots enter the final answer set is decided separately by slot-level evidence-local admission (Section 3.2).

2.3 Factuality Evaluation for Grounded Generation

The ALCE benchmark (Gao et al., 2023) reports answer-set F1 and citation quality for grounded generation tasks including QAMPARI (Amouyal et al., 2022), ASQA (Stelmakh et al., 2022), and ELI5 (Fan et al., 2019). We use QAMPARI as the headline task because its output consists of discrete entity slots, making slot-level final-inclusion decisions directly observable. Applying our view to long-form tasks such as ASQA and ELI5 would require first defining larger units, such as claims or passages, and then deciding which units should enter the output.

Our admission rule is implemented with factuality verifiers such as MiniCheck (Tang et al., 2024) and WeCheck (Wu et al., 2023), which score whether a proposed slot is supported by local evidence. We also compare against holistic factuality scorers such as AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022), which score complete outputs rather than individual answer slots. Related systems such as Self-RAG (Asai et al., 2024) and PAVE (Huang et al., 2026) integrate verification or validation into the generation pipeline. Our focus is different: we ask where the final inclusion decision should be made for answer-set generation, and test whether the resulting gains depend on matched retrieved evidence.

Algorithm 1 Proposal-Admission Decoding

Require: query q, retrieved evidence E, baseline decoder D_E, proposer decoder D_P, verifier V, threshold τ, selective-repair flag selective

$Y_0 \leftarrow \text{slots}(D_E(q, E))$
if selective **and** $Y_0 \neq \emptyset$ **then**
  return Y_0 ▷ skip proposer on populated baselines
end if
$C \leftarrow \text{slots}(D_P(q, E)) \setminus Y_0$
$Y \leftarrow Y_0$
for $s \in C$ **do**
  $g_s \leftarrow V(F(E, s),\ h(q, s))$ ▷ slot-local NLI score
  if $g_s \geq \tau$ **then**
    $Y \leftarrow Y \cup \{s\}$
  end if
end for
return Y

3 Proposal-Admission Decoding

We study grounded answer-set generation, where the model must output answer slots supported by retrieved evidence. RAG decoders can place evidence intervention at different granularities: *trajectory-level* interventions shift per-step decoding logits, *sequence-level* interventions score complete outputs, and *slot-level* interventions judge each candidate slot against matched evidence. These boundaries answer different questions: whether evidence should change the generation trajectory, whether a complete output is globally plausible, or whether an individual slot is supported by matched evidence. Figure 1 illustrates the Proposal-Admission Decoding interface: a conservative repair gate decides whether to invoke the contrastive proposer, the proposer supplies candidate slots, and slot-level admission decides which proposed slots enter the final answer set.

Algorithm 1 gives the decoding interface. PAD enforces two separation rules. First, the proposer surfaces candidate slots but does not decide final answer-set inclusion: a surfaced slot enters the final answer set only after admission. Second, admission only decides which proposed slots to accept; it does not search for additional candidates. These rules separate candidate discovery from evidence-grounded acceptance. The interface is independent of the particular proposer or verifier; concrete component choices are specified in Section 4.

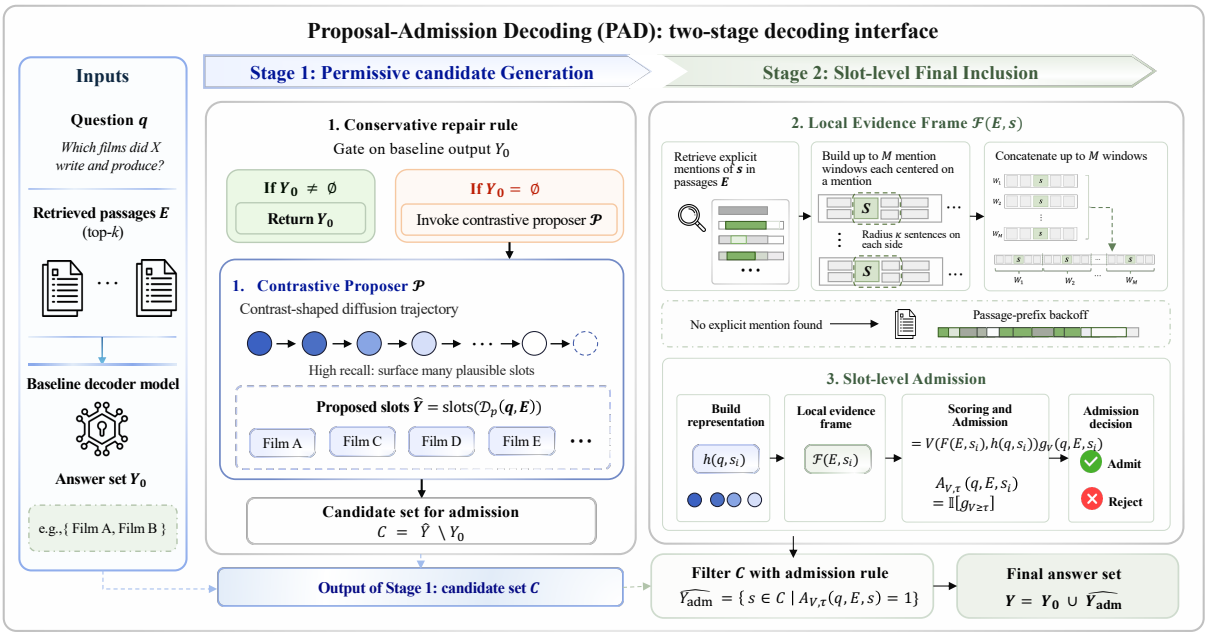

Figure 1: Overview of Proposal-Admission Decoding. In the main QAMPARI setting, the contrastive proposer is invoked only when the retrieved-evidence baseline returns an empty answer set; otherwise Y_0 is returned unchanged. The proposer supplies candidate slots $C = \text{slots}(D_P(q, E)) \setminus Y_0$, and slot-level admission adds only candidates supported by matched local evidence, yielding $Y = Y_0 \cup Y_{\text{adm}}$.

3.1 Trajectory-Level Intervention Boundary

For a query q and retrieved passages $E = \{e_i\}_{i=1}^{K}$, the system outputs a flat answer set Y. We write $\text{slots}(z)$ for the normalized, deduplicated set of answer slots parsed from an output string z. Let D_E denote the retrieved-evidence baseline decoder and define $Y_0 = \text{slots}(D_E(q, E))$.

A masked diffusion LM iterates from a fully masked $x^{(T)}$ down to $x^{(0)}$, sampling from an evidence-conditioned distribution $p_E(\cdot \mid q, E, x^{(t)})$ at each step. A direct evidence-contrast intervention computes guided logits $\ell^{(t)}_{\text{guided}} = \log p^{(t)}_E + \alpha \cdot \Delta^{(t)}_E$, with $\Delta^{(t)}_E = \log p^{(t)}_E - \log p^{(t)}_0$, applied at every denoising step $t \in \{1, \ldots, T\}$, where p_0 is the no-evidence distribution. The contrast bias propagates through the coupled denoising trajectory, and the resulting $x^{(0)}$ is treated as the final answer. CAD-D (Shi et al., 2024; Wang et al., 2025), A-CFG (Li et al., 2025), and ARAM (Kim and Ye, 2026) differ in how they set the guidance but keep this trajectory-level intervention point. This boundary is natural for applying evidence contrast, but it lets the same trajectory both discover candidates and decide which slots enter the final answer set.

3.2 Two-Stage Decoding Interface

We factor grounded decoding into two decisions, $P(q, E, Y_0) \to C$ and $A(q, E, s) \to \{0, 1\}$, where P is a proposer that returns candidate slots C and A is an admission decision for each slot s. The final output is $Y = Y_0 \cup \{s \in C : A(q, E, s) = 1\}$; if no proposed slot passes admission, the system preserves the retrieved-baseline answer.

Proposer. The proposer runs a permissive decoder D_P and returns the slots absent from the retrieved baseline, $P(q, E, Y_0) = \text{slots}(D_P(q, E)) \setminus Y_0$. The candidate set is obtained by normalizing and deduplicating the proposer output, then removing slots already present in Y_0. The interface supports selective repair: when Y_0 is nonempty, the proposer is skipped and the retrieved-baseline answer set is left unchanged. In our main QAMPARI setting, we run the contrastive proposer only when the baseline output is empty. The proposer expands the candidate set relative to the retrieved-evidence baseline while leaving final answer-set inclusion to admission.

Admission as the Final-Inclusion Boundary. Admission is the final-inclusion boundary over the fixed candidate set. It evaluates whether each proposed slot is supported by the matched evidence for the query, and accepted slots are added to the final answer set. Proposal-only decoding and holistic factuality scoring place evidence intervention at different boundaries of the pipeline, so they test

different forms of candidate generation and final-inclusion behavior.

3.3 Evidence-Local Admission

Slot-local evidence frame and hypothesis. For a candidate slot, we construct a short hypothesis tying the slot to the query and score it against the local retrieved evidence frame. Let $F(E, s)$ denote the local evidence frame: up to M mention windows, each built around an explicit mention of s, containing a sentence-neighborhood radius of κ on each side, and truncated to the configured per-window character budget. When no explicit mention is found, we use a passage-prefix backoff under the same per-window budget. The window radius κ controls how much local context is shown to the verifier, and the prefix backoff keeps the admission rule defined even when a candidate slot has no explicit surface mention. The query-slot hypothesis $h(q, s) = \text{verbalize}(q, s)$ uses a deterministic, query-pattern-indexed template that preserves the queried relation: for example, a query asking which films a person wrote and produced is verbalized as a claim that s is a film written and produced by that person. The full template list is in Appendix A.5; no LLM is used to generate hypotheses.

Verifier score and threshold. Following factuality scoring over claim–document pairs (Tang et al., 2024; Laban et al., 2022), we define a slot-local admission score $g_V(q, E, s) = V(F(E, s), h(q, s))$ and a thresholded decision $A_{V,\tau}(q, E, s) = \mathbb{I}[g_V(q, E, s) \geq \tau]$. The same interface can also use deterministic structural rules, e.g., $A_{\text{det}}(q, E, s) = \bigwedge_{r \in \mathcal{R}} r(q, F(E, s), s)$, where $\mathcal{R}$ contains the structural checks specified in Appendix B.

4 Experimental Setup

Task, evidence, and backbone. We evaluate on ALCE-QAMPARI full-dev with N=1000 examples (Gao et al., 2023; Amouyal et al., 2022). Each output is a flat answer set, so each generated entity slot can be scored by gold overlap and checked against retrieved evidence. All methods use the same ALCE-provided GTR top-5 retrieved passages for each query and are paired against the same retrieved-evidence baseline output $Y_0 = \text{slots}(D_E(q, E))$. The masked diffusion backbone is LLaDA-8B-Instruct with 64 denoising steps. The retrieved-evidence baseline has answer-set F1 of 8.28 (top-5 F1 12.85) and produces an empty Y_0 on 11.6% of examples; the selective-repair proposer therefore operates on a small but well-defined slice while preserving populated baselines.

Operating point and boundaries compared. We compare trajectory-level final-output interventions, proposal-only decoding, sequence-level holistic filtering, and slot-level evidence-local admission. In the main QAMPARI setting, we use a conservative repair rule: D_P is invoked only when Y_0 contains no answer slots. A non-empty ablation tests the same proposal-admission boundary under more aggressive proposal routing (Appendix C.2). Trajectory-level rows include no-evidence decoding, uniform CAD-D (Shi et al., 2024; Wang et al., 2025), A-CFG-style confidence guidance (Li et al., 2025), ARAM-style SNR guidance (Kim and Ye, 2026), and raw SPREAD-style relevance (Yu et al., 2026); hyperparameters are fixed before evaluation and listed in Appendix A.6. Proposal-admission rows use SPREAD-style relevance as the candidate source with $C = \text{slots}(D_P(q, E)) \setminus Y_0$. This proposer uses the same backbone and retrieved evidence as the baseline, with no additional proposer training or retrieval changes.

Admission rules. The slot-level local admission rules use a sentence-neighborhood radius κ=2, at most M=4 mention windows, and a 1400-character cap per window, with MiniCheck-Flan (Tang et al., 2024) at τ=0.25 and WeCheck (Wu et al., 2023) at τ=0.60. A deterministic local-binding reference is reported as a conservative comparison. Two holistic factuality scorers, AlignScore-large (Zha et al., 2023) at τ=0.25 and SummaCConv (Laban et al., 2022) at τ=0.05, serve as sequence-level comparisons. For these holistic comparisons, the scorer evaluates the complete candidate-augmented output against the retrieved evidence. If the sequence-level score exceeds the threshold, the proposed output is accepted as a whole; otherwise the system falls back to the retrieved-evidence baseline output. Thresholds are selected by the matched-evidence dependence criterion rather than by real-evidence F1 alone. For each scorer, we consider thresholds with positive real-evidence ΔF1 and report the operating point with the lowest shuffled-evidence ΔF1 among them. This reflects the grounding precondition used throughout the paper: a reported gain should depend on matched evidence rather than persist under evidence mismatch. Full sweeps are

in Appendix C.10. Appendix C.4 further reports calibration–holdout splits to check that the selected results are not artifacts of threshold choice alone.

Evidence controls and metrics. For admission methods, the shuffled-evidence test fixes the candidate set C_i and replaces only the evidence supplied to admission with $E_{\pi(i)}$ under a permutation π:

$$Y_i^{\text{shuf}} = Y_{0,i} \cup \{s \in C_i : A(q_i, E_{\pi(i)}, s) = 1\}.$$

The proposer and candidate set are fixed; only the evidence supplied to admission changes. Scoring Y_i^{shuf} against the same gold answers gives Δ_{shuf}; if an admission gain depends on matched evidence, it should disappear under this mismatch. Proposal-only decoding has no separate admission boundary and so is not assigned a shuffled-admission score. We summarize matched-evidence and shuffled-evidence deltas by the binding ratio

$$\rho_{\text{bind}}(m) = \frac{\Delta_{\text{shuf}}(m)}{\max(\Delta_{\text{real}}(m),\ \varepsilon)},$$

reported for positive-gain results, with ε a small constant to prevent division by zero. For results with positive real-evidence gains, values near 0 indicate that the gain collapses under evidence mismatch, while values near 1 indicate that the gain largely survives evidence mismatch. Slot-binding counterfactuals additionally swap the evidence, query, slot surface, and a nearby hard-negative slot to test whether admission tracks the intended query-slot relation rather than mere mention or plausibility (Appendix D). We also use a mention-matched negative evidence control that preserves the slot mention by importing passages from other examples (Appendix C.5).

Our primary public metric is official ALCE answer-set ΔF1, paired against the retrieved baseline following paired evaluation practice (Dror et al., 2018). For grounded answer-set generation, an F1 gain is interpretable as grounded only if it depends on matched evidence. The main diagnostic tuple is (ΔF1, W/L/T, Δ_{shuf}, ρ_{bind}); we additionally track true output changes for changed-case analysis and ablations. We interpret official-F1 gains together with Δ_{shuf} and ρ_{bind}: a gain that survives evidence shuffling is treated as evidence-leaky rather than evidence-bound admission. Unless explicitly labeled $\Delta\text{F1}_{\text{top5}}$, all main-text deltas are official ALCE answer-set F1; top-5 F1 is reported alongside official F1 only in appendix ablations. Confidence intervals are descriptive bootstrap 95% intervals over paired example deltas, and sign tests over non-tied examples are reported in the same descriptive role for the selected results.

Intervention	ΔF1	95% CI	W/L/T
CAD-D	-0.184	$[-0.676, +0.312]$	71/112/817
A-CFG	-0.335	$[-0.658, -0.022]$	52/78/870
ARAM	-0.825	$[-1.469, -0.195]$	124/165/711
SPREAD	$+0.180$	$[-0.418, +0.785]$	116/157/727

Table 1: Direct trajectory-level interventions as final answer-set boundaries on full-dev QAMPARI. ΔF1 is official ALCE answer-set F1 against the retrieved-evidence baseline; intervals are bootstrap 95% CIs over paired example deltas. These rows motivate a separate slot-level final-inclusion boundary.

Because the main setting only repairs empty baseline outputs, the admission verifier is called on 282 candidate pairs in total, averaging 0.28 verifier calls per example; detailed cost statistics are reported in Appendix A.6.

5 Where Evidence Should Intervene

We test whether evidence intervention is reliable at each of the granularities introduced in Section 3. The primary outcome is official ALCE answer-set F1, paired against the retrieved-evidence baseline with the same LLaDA-8B-Instruct decoder; we use the evidence-dependence diagnostics defined in Section 4. Output changes and changed-case inspection are diagnostics: they explain the public gains, but do not replace the public metric. The question is therefore not only whether F1 improves, but whether the improvement disappears when the admission evidence is mismatched. We therefore read each operating point by both its public-metric gain and its evidence-dependence signature.

5.1 Direct Contrast Is Unstable as a Final-Inclusion Boundary

The first test is whether trajectory-level contrast can serve as the final answer-set boundary. Table 1 shows that trajectory-level contrast is not a reliable final-inclusion boundary in this setting. Uniform CAD-D, A-CFG-style guidance, and ARAM-style guidance all lose official F1 as final trajectory-level decision rules. Raw SPREAD-style relevance is the informative exception: it is positive on average, but it has a bootstrap interval crossing zero, 273 score-changing non-ties, 831 true output changes, and 157 paired losses. These results encapsulate the boundary mismatch: contrast can expose missing candidates, but the contrast-shaped trajectory is too unstable to serve as a reliable final-inclusion

decision.

5.2 Proposal-Only Decoding Supplies Candidate Discovery

Table 2 separates three phenomena that official F1 alone conflates: candidate discovery, global plausibility filtering, and matched-evidence-bound final inclusion. It therefore focuses on what happens after the boundary is moved: contrast supplies proposals, and different admission rules decide final answer-set inclusion.

Proposal-only decoding is informative because it isolates candidate discovery from admission. The empty-only SPREAD proposal uses SPREAD candidates only when the retrieved baseline produces no answer slots, and reaches $+0.539$ F1 with no paired losses and 83 output changes. It shows that contrastive trajectories can surface candidate slots, but without admission every proposed slot enters the final answer set by construction; plausible-but-unbound slots are added alongside bound ones. Thus proposal-only gains demonstrate candidate discovery, while matched-evidence dependence requires a separate final-inclusion boundary. This row should therefore be read as a candidate-discovery control rather than as a grounded admission method.

5.3 Holistic Scorers Retain Gains Under Shuffled Evidence

If ranked by official F1 alone, holistic factuality scorers would look competitive or even superior. AlignScore-large improves by $+0.424$ with real evidence; SummaCConv reaches $+0.542$. The shuffled-evidence control reverses this interpretation. AlignScore-large retains $+0.453$ under shuffled evidence, slightly higher than the real-evidence delta; SummaCConv retains the same $+0.542$ in both conditions. Holistic scorers are effective global plausibility filters, but under this fixed-candidate control their gains are not tied to the matched retrieved evidence. This does not mean that holistic scorers are ineffective; rather, it shows that they instantiate a different intervention boundary from slot-level evidence-local admission. Figure 2 visualizes this boundary separation on the real-evidence versus shuffled-evidence gain plane.

5.4 Slot-Level Admission Produces Matched-Evidence-Dependent Gains

At the slot-level final-inclusion boundary, Proposal-Admission Decoding with MiniCheck-Flan admission at threshold 0.25 improves F1 by $+0.356$ with a bootstrap 95% interval of $[+0.147, +0.611]$ and a two-sided sign-test $p{=}4.88 \times 10^{-4}$. Under shuffled evidence the same admission rule collapses to zero gain with no output changes and no admitted slots. Proposal-Admission Decoding with WeCheck at threshold 0.60 gives an independent slot-level point at $+0.360$ with a 95% interval of $[+0.161, +0.609]$ and sign-test $p{=}2.44 \times 10^{-4}$, and also collapses under shuffled evidence with zero output changes and zero admitted slots.

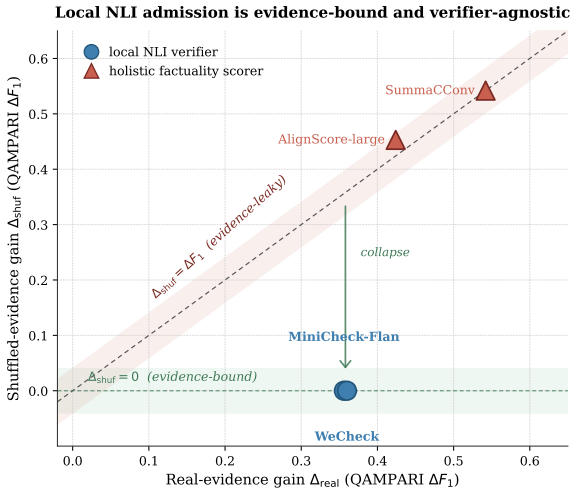

Figure 2: Evidence-dependence plane for the admission rows in Table 2. Slot-level admission collapses to $\Delta_{\text{shuf}}{=}0$, while holistic scorers stay near the leaky diagonal $\Delta_{\text{shuf}}{=}\Delta_{\text{real}}$.

The gain is selective by design: admission changes final answer-set inclusion only when the proposed slot passes a local matched-evidence test. This is the expected signature of an evidence-bound boundary, and it contrasts with proposal-only decoding and evidence-leaky holistic scoring. Appendix C.11 shows that two additional MiniCheck variants land in the same positive, no-paired-loss operating region, indicating that the evidence-bound behavior is not specific to a single verifier checkpoint.

The conservative repair setting is a precision-oriented choice for the main experiment, not a limitation of the interface. The non-empty routing ablation addresses whether the boundary principle only holds on the empty-baseline slice. When the proposer is allowed to add to populated baselines, MiniCheck-Flan admission reaches $+0.684$ F1 at $\tau{=}0.25$, but paired losses appear ($49/87/864$ W/L/T). The shuffled-evidence delta remains essentially zero (-0.002), so the evidence-

Operating point	Boundary	ΔF1	W/L/T	Shuf. Δ	ρ_{bind}
Trajectory-level final boundary					
SPREAD final answer (Yu et al., 2026)	trajectory	+0.180	116/157/727	–	–
Proposal-only decoding (no admission test)					
Proposal-only	proposal-only	+0.539	18/0/982	–	–
Sequence-level holistic filtering					
Holistic (AlignScore) (Zha et al., 2023)	sequence	+0.424	14/0/986	+0.453	1.07
Holistic (SummaCConv) (Laban et al., 2022)	sequence	+0.542	18/0/982	+0.542	1.00
Slot-level evidence-local admission					
PAD (MiniCheck-Flan) (Tang et al., 2024)	**slot-level**	**+0.356**	**12/0/988**	**0.000**	**0.00**
PAD (WeCheck) (Wu et al., 2023)	slot-level	+0.360	13/0/987	0.000	0.00
PAD (deterministic)	slot-level	+0.292	9/0/991	–	–

Table 2: QAMPARI results organized by intervention boundary. PAD denotes Proposal-Admission Decoding. All rows are paired against the same retrieved-evidence baseline with LLaDA-8B-Instruct. ΔF1 is official ALCE answer-set F1 against that baseline; Shuf. Δ is the same delta with admission evidence permuted across queries; $\rho_{\text{bind}}=\Delta_{\text{shuf}}/\Delta_{\text{real}}$. Dashes mark rows for which shuffled-evidence deltas are not reported. PAD-style slot-level admission shows matched-evidence dependence: $\rho_{\text{bind}}\approx 0$ for slot-level rows, compared with $\rho_{\text{bind}}\approx 1$ for holistic rows with strong ΔF1 gains.

Routing	τ	ΔF1	W/L/T	Shuf. Δ / ρ_{bind}
Selective repair	0.25	+0.356	12/0/988	0.000 / 0.00
Non-empty routing	0.25	+0.684	49/87/864	−0.002 / ≈ 0
Non-empty routing	0.40	+0.634	47/83/870	– / –

Table 3: Routing coverage changes the precision–recall operating point for MiniCheck-Flan admission. Selective repair is conservative and high-precision; non-empty routing increases mean F1 while introducing paired losses. The near-zero shuffled delta at $\tau=0.25$ shows that matched-evidence dependence is preserved beyond the conservative repair setting.

dependence property survives the higher-coverage setting while precision is traded for recall (Table 3; Appendix C.2).

5.5 Stress Tests and Cross-Task Check

We run additional stress tests to check that the headline result is not a narrow artifact. First, the result is stable across admission thresholds and local-window sizes. MiniCheck-Flan remains positive from threshold 0.10 to 0.45, admissions under shuffled evidence vanish for thresholds at or above 0.10, and $\kappa \in \{1, 2, 3\}$ gives the same $12/0/988$ W/L/T pattern (Appendices C.10 and C.3). Calibration–holdout splits remain positive and loss-free, reducing the concern that the threshold is selected on the reporting set (Appendix C.4). Mention-based controls show that the gains are not driven by prefix backoff or mere mention availability: 26 of 28 changed examples are mention-window driven, and a mention-matched negative control admits 0 of 32 relation-mismatched cases (Appendices C.1 and C.5). This stricter control suggests that the admission rule is not merely detecting slot mentions, but checking whether the local evidence supports the queried relation. Together, the coverage ablation and stress tests support the interpretation that the observed gains come from the slot-level admission boundary rather than from the conservative repair setting alone. Additional qualitative cases and slot-binding counterfactuals further localize the remaining failure mode: at the operating thresholds, both local verifiers reject evidence-swap and query-swap perturbations, while near-miss slot swaps and hard negatives remain challenging (Appendix D).

As a lightweight cross-task check, a span-set repair variant on MultiSpanQA improves exact F1 by +0.805, while wrong-evidence controls collapse the unguarded gate but leave final guarded outputs unchanged (Appendix C.8).

6 Conclusion

Proposal-Admission Decoding separates contrastive candidate generation from slot-level evidence-local admission, using contrastive trajectories to surface candidates rather than to decide final answer-set inclusion. On ALCE-QAMPARI, proposal-only decoding shows candidate discovery, holistic factuality scorers retain gains under shuffled evidence, and slot-level admission yields gains that collapse under shuffled evidence. These results support a boundary principle: candidate generation and final inclusion should be separated, and grounded decoders should be evaluated by both official F1 and matched-evidence dependence.

Limitations

We focus on flat answer-set generation because it provides a clean unit for testing intervention boundaries. Extending the framework to claim-, tuple-, citation-, or passage-level admission requires redefining the local evidence frame $F(E, \cdot)$ and hypothesis $h(q, \cdot)$ for the new unit; the binding-ratio diagnostic remains applicable once that unit is defined. All headline experiments use LLaDA-8B-Instruct because coupled multi-step denoising makes the trajectory-level boundary mismatch especially visible, though the MultiSpanQA check (Section 5.5) suggests the boundary view remains meaningful for span sets.

The headline operating point gates the proposer on $Y_0{=}\emptyset$ to prioritize precision and preserve populated retrieved-baseline answers. The non-empty ablation (Appendix C.2) raises mean ΔF1 but introduces paired losses, reflecting a precision–recall tradeoff under broader routing; its near-zero ρ_{bind} suggests that matched-evidence dependence extends beyond the conservative repair setting. Finally, the slot-swap and hard-negative failures in Appendix D identify a remaining admission bottleneck: off-the-shelf NLI models remain weak at fine-grained discrimination against near-miss slot surfaces.

References

Samuel Joseph Amouyal, Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. 2022. QAMPARI: an open-domain question answering benchmark for questions with many answers from multiple paragraphs. *CoRR*, abs/2205.12665.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Giulio Corallo and Paolo Papotti. 2026. Parallel context-of-experts decoding for retrieval augmented generation. *CoRR*, abs/2601.08670.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Tianyi Huang, Caden Yang, Emily Yin, Eric Wang, and Michael Zhang. 2026. PAVE: Premise-aware validation and editing for retrieval-augmented LLMs. *CoRR*, abs/2603.20673.

Jaemin Kim and Jong Chul Ye. 2026. Adaptive guidance for retrieval-augmented masked diffusion models. *CoRR*, abs/2603.17677.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 2421–2431. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163–177.

Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. 2025. Adaptive classifier-free guidance via dynamic low-confidence masking. *CoRR*, abs/2505.20199.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *CoRR*, abs/2502.09992.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 130136–130184. Curran Associates, Inc.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8818–8847. Association for Computational Linguistics.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025, Albuquerque, New Mexico, April 29 - May 4, 2025*, pages 11636–11652. Association for Computational Linguistics.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023. Wecheck: Strong factual consistency checker via weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 307–321. Association for Computational Linguistics.

Chuanyue Yu, Jiahui Wang, Yuhan Li, Heng Chang, Ge Lan, Qingyun Sun, Jia Li, Jianxin Li, and Ziwei Zhang. 2026. Unlocking the potentials of retrieval-augmented generation for diffusion language models. *CoRR*, abs/2601.11342.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

A Reproducibility Details

This appendix lists the fixed components needed to reproduce the main QAMPARI experiments: retrieval, backbone decoding, proposal construction, admission rules, normalization, scoring, and evidence controls. All paired comparisons use the same retrieved-evidence baseline output Y_0 and the same retrieved passages.

A.1 Dataset and Model

The main experiment uses ALCE-QAMPARI full-dev with N=1,000 examples and the ALCE-provided GTR top-5 retrieved passages. The backbone is LLaDA-8B-Instruct with 64 denoising steps, 64 maximum generated tokens, and block length 32. The retrieved-evidence baseline has answer-set F1 8.28 (top-5 F1 12.85) and produces an empty answer set on $116/1000$ examples (11.6%). QAMPARI is the headline task because each output is a flat answer set, which makes slot-level admission directly checkable against local evidence.

A.2 Decoding and Proposal Sources

The retrieved-evidence baseline decodes with the retrieved passages. The no-evidence baseline removes retrieved passages. Uniform CAD-D contrasts evidence-conditioned and no-evidence trajectories uniformly. The A-CFG-style and ARAM-style rows use confidence- and SNR-inspired guidance variants. For direct-contrast rows, SPREAD-style relevance is evaluated as a final trajectory-level boundary. For proposal-admission rows, the same SPREAD-style decoder is used only to supply candidate slots; final answer-set inclusion is decided by admission.

A.3 Normalization and Scoring

We use deterministic surface normalization before slot comparison: case folding, punctuation normalization, whitespace normalization, and dash normalization. We report official ALCE answer-set F1 and paired W/L/T against the retrieved-evidence baseline. When a method changes the output but produces zero metric delta, we record it as a changed-zero case rather than treating the tie as behaviorally neutral.

A.4 Main Settings

Setting	Value
Backbone	LLaDA-8B-Instruct
Denoising steps	64
Maximum generated tokens	64
Block length	32
Dataset	ALCE-QAMPARI full-dev, $N = 1000$
Retrieved evidence	ALCE-provided GTR top-5 passages
Reference baseline	Retrieved-evidence decoding
Baseline empty examples	$116/1000$ (11.6%)
Candidate proposal	SPREAD-style contrastive proposal
Proposal gate	selective repair, $Y_0=\emptyset$
Admission frame	sentence-neighborhood radius $\kappa=2$, $M=4$ mention windows, 1400 char cap
Canonical local admission	MiniCheck-Flan, threshold 0.25
Independent local admission	WeCheck, threshold 0.60
Deterministic local reference	explicit mention plus structural query binding
Normalization	Case, punctuation, whitespace, dash normalization
Evidence tests	Shuffled evidence and slot-binding counterfactuals

Table 4: Main settings used for the QAMPARI experiments.

A.5 Query-Slot Verbalization Templates

The local admission rules score a deterministic claim $h(q, s)$ against a local evidence frame. Before template matching, the question is stripped of a terminal question mark and internal whitespace is normalized. The closed answer-type list is {`movie`, `film`, `book`, `song`, `album`, `manga`, `language`, `school`, `college`, `university`, `city`, `country`}. Table 5 lists the complete template inventory used for QAMPARI; no LLM is used to generate hypotheses.

Question pattern	Hypothesis $h(q, s)$
`for which X did U R`	`U R for s.`
`who are/is/was/were R`	`s is R.`
`what A was/were/is/are R`, with A in a closed list of answer types	`s was R.`
`what A was/were/is/are R`, otherwise	`s is A that was R.`
`which A was/were/is/are R`	`s is A that was R.`
Fallback	`s is a correct answer to the question: q.`

Table 5: Deterministic query-slot verbalization templates. Here s is the proposed slot, q is the normalized question, A is the answer-type phrase, U is the query subject, and R is the remaining queried relation phrase.

A.6 Fixed Hyperparameters

Table 6 lists the fixed settings used by the trajectory-level comparisons, proposal source, and admission rules. All values are fixed before the full-dev evaluation.

Component	Fixed setting
Backbone decoding	LLaDA-8B-Instruct, 64 denoising steps, 64 maximum generated tokens, block length 32
Retrieved baseline	evidence-conditioned decoding, no contrast
No-evidence baseline	same decoding, retrieved passages removed
Uniform CAD-D	`cad_d`, uniform schedule, $\alpha=0.25$
A-CFG-style guidance	`a_cfg_confidence`, guidance scale $w=0.5$, remask ratio $\rho=0.7$, start progress 0.0
ARAM-style guidance	`aram_snr_guidance`, $\lambda=1$, $\beta=0.1$, $\epsilon=10^{-6}$
SPREAD-style trajectory/proposer	`spread_query_relevance` with the same retrieved passages and backbone; no added proposer training or retrieval changes
Prompt evidence budget	top-5 retrieved passages, 260 characters per passage
Selective repair gate	run the proposer iff $Y_0 = \emptyset$ ($116/1000$ examples)
Local evidence frame	sentence-neighborhood radius $\kappa=2$, at most $M=4$ mention windows, max 1400 characters per window, passage-prefix backoff if no mention is found
MiniCheck-Flan admission	threshold $\tau=0.25$
WeCheck admission	threshold $\tau=0.60$
AlignScore-large holistic scorer	threshold $\tau=0.25$
SummaCConv holistic scorer	threshold $\tau=0.05$

Table 6: Fixed hyperparameters for the main QAMPARI comparisons and admission rules.

Cost accounting. Proposal-Admission Decoding adds one SPREAD-style decoding pass and one verifier pass on top of the retrieved-evidence baseline when the selective repair gate is active. Because the main setting only repairs empty baseline outputs, both extra passes are skipped on $884/1000$ examples (88.4%), where the baseline already produces a non-empty answer set. Across the 1,000-example run, the admission verifier evaluates 282 candidate (slot, query) pairs in total, averaging 0.28 verifier calls per example and 2.43 calls per routed example among the 116 examples where the proposer runs. We report candidate counts as the primary cost summary because wall-clock time depends heavily on implementation and hardware; the SPREAD proposer dominates the added computation relative to the verifier.

B Deterministic Local-Admission Reference

B.1 Check Inventory

The deterministic local-admission reference retains an explicit-mention precondition plus six query-level binding obligations. Table 7 lists the checks

Check	Used?	Obligation
Mention precondition	yes	Candidate slot is explicitly mentioned in evidence.
Query-copy veto	no	Broad lexical query-copy exclusion.
Numeric/temporal match	yes	Numeric or temporal constraints are locally satisfied.
Entity conjunction	yes	Multi-entity conjunction is locally supported.
Role conjunction	yes	Role conjunctions are locally supported.
Answer type	yes	Candidate matches expected answer type.
Queried relation	yes	Local frame contains the queried relation.
Query subject	yes	Local frame binds the query subject.

Table 7: Deterministic local-admission checks. The broad query-copy veto is dropped; the remaining checks are retained as a conservative structural reference.

used by the reference.

B.2 Deterministic Implementation

The deterministic reference does not call an NLI model, an LLM verifier, or a dependency parser. It applies fixed surface procedures to the selected local evidence frame: case-folded string normalization, regular expressions for temporal expressions and conjunction patterns, capitalized-span extraction for query entities, hand-written answer-type cues, and a fixed relation-lemma lexicon for roles such as writing, directing, producing, composing, attending, and representing. The mention precondition requires exact normalized slot overlap in the frame. Numeric and temporal checks require explicit year overlap satisfying the query comparator, e.g., *after 2010* requires a year greater than 2010 in the slot-local context. Entity and role conjunction checks require both sides of the detected conjunction to appear within the same slot-local context. The queried-relation check uses a tighter window around the slot mention and requires at least one relation lemma from the query's relation group. These rules are deliberately surface-bound: they provide a reproducible structural reference, not a learned semantic verifier.

B.3 Admission Procedure

Operationally, the deterministic reference takes a question q, retrieved evidence E, a retrieved-baseline answer set Y_0, and contrastive candidate slots C. It removes slots already present in Y_0, locates each remaining slot in E, evaluates the retained checks in the local evidence frame, and adds admitted slots to Y_0. If no slot passes, the system preserves the retrieved-baseline answer.

1. Extract candidate answer slots from the contrastive proposal and remove slots already present in the baseline.
2. For each slot, locate the closest retrieved evidence frame that explicitly mentions the slot.
3. Reject the slot if the mention precondition fails.
4. Evaluate the six retained query-level obligations in the local slot frame.
5. Admit the slot only if all applicable obligations pass.
6. Add admitted slots to the retrieved-baseline answer set; if no slot passes, preserve the retrieved-baseline answer.

B.4 Mention Precondition vs. Query-Copy Veto

The mention precondition and query-copy veto look superficially similar because both refer to lexical overlap, but they play opposite roles. The mention precondition is a grounding requirement: a slot absent from retrieved evidence cannot be cited as evidence-supported. The query-copy veto is a broad lexical exclusion: it rejects a slot because it overlaps with the question. The Dan Gilroy case in Appendix D.3 supports the mention precondition and illustrates the failure mode of the broad veto. Multi-role self-answers, such as Dan Gilroy directing a film he wrote, are valid even though the answer entity appears in the query. They show why admission should rely on structural relation binding rather than shallow lexical exclusion.

C Additional Results

C.1 Mention-Window vs. Backoff Admission

We test whether the admission gain is driven by the prefix backoff rather than by slot-local evidence frames. Across the 282 candidate slot pairs evaluated by MiniCheck-Flan at τ=0.25 on QAMPARI full-dev, 246 (87.2%) had at least one mention-window in the retrieved evidence and 36 (12.8%) backed off to the passage-prefix fallback. The admission rate is comparable on the two frame types (21.1% on mention candidates, 27.8% on backoff candidates), and of the 62 total admitted slots,

Setting	Official ΔF1	$\Delta F1_{top5}$	W/L/T	ρ_{bind}
Selective repair (τ=0.25)	+0.356	+0.485	12/0/988	0.00
Non-empty (τ=0.10)	+0.655	+0.686	53/107/840	–
Non-empty (τ=0.25)	+0.684	+0.739	49/87/864	≈ 0
Non-empty (τ=0.40)	+0.634	+0.697	47/83/870	–

Table 8: Non-empty proposal ablation for MiniCheck-Flan admission. When the proposer can add to populated baselines, mean F1 increases but paired losses appear (49 wins vs. 87 losses at τ=0.25). The shuffled-evidence official ΔF1 at τ=0.25 is −0.002 on the same examples, giving ρ_{bind}≈0: matched-evidence dependence remains, but precision is traded for recall.

κ	Official ΔF1	$\Delta F1_{top5}$	W/L/T
1	+0.336	+0.468	12/0/988
2 (canonical)	+0.356	+0.485	12/0/988
3	+0.351	+0.476	12/0/988

Table 9: Window-size sensitivity at τ=0.25. The 12 winning examples and zero losses are identical across $\kappa \in \{1, 2, 3\}$; official ΔF1 varies only between +0.336 and +0.356. The canonical operating point is not tied to a single κ.

52 (83.9%) come from mention-windows and 10 (16.1%) come from backoff frames.

At the example level, the concentration in mention-based frames is clearer. Of the 28 output-changing examples for MiniCheck-Flan, 26 (92.9%) are mention-only (every admitted slot in the example came from a mention-window), 0 are backoff-only, and 2 are mixed (at least one mention admit plus at least one backoff admit). Backoff frames therefore complement mention admits in a small number of changed examples but never solely produce a change.

Under shuffled evidence the dependence behavior is consistent: only 4 of the 282 candidates retain any mention in the shuffled passages (since the slot rarely appears in a different example's evidence), and MiniCheck-Flan admits 0 from either group. Backoff is therefore a minority contribution that does not by itself drive the matched-evidence gain.

C.2 Non-Empty Proposal Ablation

The canonical operating point in Table 2 is selective repair: it routes the proposer only when $Y_0 = \emptyset$. The non-empty ablation removes this gate, allowing the SPREAD-style proposer to add slots to populated baselines as well (Table 8). The sweep uses the same MiniCheck-Flan admission scores and varies only the threshold, so it tests whether stricter admission can compensate for more aggressive proposal routing.

Two findings follow. First, selective repair is the high-precision operating point: allowing proposal on populated baselines increases mean ΔF1 but creates more paired losses than wins. These losses are not removed by simply raising the threshold: even at τ=0.40, the non-empty setting has 47 wins and 83 losses. The issue is therefore the candidate distribution exposed by populated baselines, not only the selected threshold. Second, matched-evidence dependence extends beyond selective repair: even under the more aggressive non-empty setting, the shuffled-evidence delta is essentially zero (ρ_{bind}≈0). The headline operating point reported in Section 5 is therefore the precision-oriented point in this ablation; matched-evidence dependence also holds when proposal routing is more aggressive.

C.3 Window-Size (κ) Sensitivity

The local evidence frame uses a sentence-neighborhood radius of κ around mentions; in the main experiments κ=2. Table 9 reports the same MiniCheck-Flan admission rule at τ=0.25 for $\kappa \in \{1, 2, 3\}$.

The 12/0/988 paired pattern is preserved across the three settings, and the official ΔF1 range $[+0.336, +0.356]$ is contained within a narrow band. The operating point is not sensitive to the chosen κ in this range.

C.4 Calibration–Holdout Threshold Split

To check that the reported intervals are not determined by selecting the threshold on the same examples used for reporting, we partition the 1,000 full-dev examples into a 200-example calibration set and an 800-example holdout, repeating with five random seeds. At τ=0.25 the holdout official ΔF1 is positive across all five seeds (seed 1: +0.353, 9/0/791; seed 2: +0.390, 11/0/789; seed 3: +0.276, 9/0/791; seed 7: +0.445, 12/0/788; seed 42: +0.431, 11/0/789). Holdout W/L/T is consistent with the full-set 12/0/988 pattern: in each split the holdout contains the bulk of the wins (with 0 losses), and any wins falling in the calibration set are reported there. The official ΔF1 ranges and the loss-free property therefore reflect the underlying operating point rather than threshold-selection on the same examples used for reporting.

C.5 Mention-Matched Negative Evidence Control

A sharper test than shuffled evidence is to keep the mention of the slot but break the query-slot relation binding. For each of the 62 slots admitted by MiniCheck-Flan at τ=0.25 on QAMPARI full-dev, we search the union of all retrieved passages across the 1,000 examples for a passage that contains the slot mention but comes from a *different* example. Such a passage preserves slot mention while replacing the original query context with one for which the slot is not a correct answer.

For 32 of the 62 admitted slots, at least one such mention-containing passage exists in another example's retrieved context (the remaining 30 slots are unique enough that no other example's retrieval mentions them). We then re-score the same query-slot hypothesis against the adversarial passage and apply the same τ=0.25 admission threshold.

Across the 32 admitted slots with an adversarial passage available, the admission rate under mention-matched negative evidence drops to $0/32$ (0%), compared to 100% under matched evidence by construction. The verifier's score is therefore not driven by the surface presence of the slot in some retrieved passage; it is driven by whether the locally retrieved evidence binds the slot to the queried relation. This complements the random shuffled-evidence collapse (where the slot rarely appears at all) by showing that even when the surface mention is preserved, an admission that is not bound to the matched query context is rejected.

C.6 A Self-RAG Reference Point

For external reference, not as part of the controlled boundary comparison, we run Self-RAG (Asai et al., 2024) (`selfrag_llama2_7b`) on full-dev QAMPARI with the same retrieved passages used by the LLaDA-8B-Instruct retrieved baseline. Against the retrieved-evidence baseline, Self-RAG yields a negative paired F1 delta: mean -3.55 (95% CI $[-4.53, -2.56]$, W/L/T $120/273/607$), and Self-RAG's mean answer-set F1 is 4.76. This is not a controlled boundary comparison: Self-RAG decodes its own reflection-token sequence and is trained for a different output style. We include it only as a reference point indicating that the LLaDA-8B-Instruct retrieved baseline used throughout this paper is non-trivial.

C.7 Deterministic Reference Variants

We evaluate three deterministic local-admission operating points: a strict full-check variant, the structural reference used in the main table, and a relation-relaxed variant. The comparison is useful because it separates metric recall from local relation binding. Relaxing the local relation check increases official F1 but weakens evidence binding, which is why that variant is not used as the structural reference.

C.8 Cross-Task Scope Checks

We ran cross-task scope checks on MultiSpanQA, BioASQ, and FanOutQA. These checks are not main-table method comparisons. MultiSpanQA is the closest secondary task because it also has span-set answers. Unlike QAMPARI's entity-centric answer sets, MultiSpanQA scores sets of contiguous answer spans by exact string matching. The proposal-admission interface transfers at the level of the admission unit: spans replace entities as the discrete candidates checked against local evidence frames. On the full validation set ($n = 653$), the span-set repair operating point improves official exact F1 from 53.481 to 54.286. A per-example paired analysis gives a mean exact-F1 delta of +0.566 with 25/4/624 W/L/T, bootstrap 95% CI $[+0.177, +0.972]$, and a two-sided sign-test $p = 1.04 \times 10^{-4}$ over non-ties. BioASQ is approximately neutral or slightly negative under repair, and FanOutQA requires pair-structured checks rather than flat slot checks. The negative and near-neutral transfer is consistent with the paper's scope claim: the admission unit must match answer structure.

The most informative cross-task signal is therefore the wrong-evidence test rather than the transfer number alone. On MultiSpanQA, the unfiltered proposal output under random evidence collapses exact-F1 from 53.48 to 1.82, while the guarded output remains at 53.48 with 0/0/653 W/L/T. Under foil evidence, the unfiltered proposal output reaches only 1.77 exact-F1, while the guarded output remains at 53.44 with 0/2/651 W/L/T. This supports the claim that admission is an evidence boundary rather than a generic plausibility score, while preserving the paper's narrow scope.

Setting	n	Exact F1	Δ	W/L/T	95% CI	Overlap Δ
Retrieved baseline	653	53.481	–	–	–	–
Span-set repair	653	54.286	+0.805	25/4/624	[+0.177,+0.972]	+0.368
Random-evidence unfiltered proposal	653	1.824	-51.657	–	–	–
Random-evidence guarded output	653	53.481	0.000	0/0/653	–	–
Foil-evidence unfiltered proposal	653	1.765	-51.716	–	–	–
Foil-evidence guarded output	653	53.439	-0.042	0/2/651	–	–

Table 10: MultiSpanQA full-validation scope check. Exact F1 is the official set metric used by the MultiSpanQA scorer. The CI is a bootstrap interval over per-example exact-F1 deltas for span-set repair; W/L/T is also computed per example. Random- and foil-evidence rows are wrong-evidence controls, not method comparisons.

C.9 Backbone Transfer Scope Check (Dream-v0)

We also ran a model-transfer scope check with Dream-v0-Instruct-7B on ALCE-QAMPARI. This is not a main-table method comparison. It asks two narrower questions: whether retrieved evidence gives a stable signal for another masked diffusion backbone, and whether our current admission script can be applied unchanged. Table 11 shows that retrieved-evidence decoding strongly improves over no-evidence decoding at $n = 1{,}000$ by about $+4.49$ official F1 with 215/60/725 W/L/T. By contrast, treating Dream's retrieved output as a proposal source and applying the unchanged MiniCheck admission script gives only $+0.024$ F1 against no evidence, because the current script only admits into examples where the no-evidence baseline is empty. This is a scope check for applying the unchanged operating point to a second backbone: thresholds and proposal mechanisms should be re-derived per backbone, rather than treated as evidence against the proposal-admission boundary itself. The valid conclusion is limited: retrieval matters for Dream, but a full-dev Dream SPREAD-style proposal generator and a re-derived admission interface are still needed before claiming backbone transfer.

C.10 Threshold and Shuffled-Evidence Stability

Table 12 reports representative threshold points for the two local admission rules and the two holistic scorer tests. The thresholds are not used as separate methods. They show how each admission rule trades admission coverage against evidence dependence. MiniCheck-Flan remains positive over a broad operating region, and its shuffled-evidence admissions vanish at thresholds at or above 0.10. WeCheck remains positive over a similarly broad region; we use threshold 0.60 because it is the conservative point where shuffled evidence produces no output changes and no admitted slots. By contrast, AlignScore-large and SummaCConv retain high shuffled-evidence gains at their high-F1 operating points, which is why they are evidence-leaky comparisons rather than local admission rules.

C.11 Full-Dev Public-Metric Significance

Table 13 gives the public-metric statistical summary for the main positive operating points. All rows use the same full-dev ALCE-QAMPARI paired scorer. Intervals are bootstrap 95% CIs over example-level official answer-set F1 deltas, and p values are two-sided sign tests over non-tied W/L examples. The local-admission gains are small in coverage but not explained by random W/L direction: all four NLI verifiers (MiniCheck-Flan, WeCheck, MiniCheck-RoBERTa, and MiniCheck-DeBERTa) are positive, loss-free, and statistically non-random under the paired sign test.

A saturated NLI verifier baseline. We also tested a T5-XXL TRUE-NLI verifier as an additional NLI-style admission score. Under our query-slot hypothesis format, its scores were saturated: all thresholds from 0.05 to 0.90 admitted the same 282 proposal slots and changed the same 83 outputs. The representative threshold 0.25 therefore behaves like an all-pass proposal point rather than a selective local-admission point, with +0.542 official F1 and 18/0/982 W/L/T. We treat this as a diagnostic baseline showing that not every NLI verifier provides an effective admission boundary under this interface.

Cross-verifier convergence. Two additional MiniCheck variants, MiniCheck-RoBERTa and MiniCheck-DeBERTa, land in the same positive, loss-free operating region as MiniCheck-Flan and WeCheck. At the canonical threshold τ=0.25, MiniCheck-RoBERTa gives $\Delta F_1 = +0.299$ with 95% CI $[+0.107, +0.550]$ and 10/0/990 W/L/T (sign-test p=1.95×10^{-3}); MiniCheck-

Setting	n	Baseline F1	Method F1	Δ F1	W/L/T
Dream evidence vs. no evidence	1,000	2.167	6.654	+4.487	215/60/725
Retrieved-as-proposal + MiniCheck, $t = 0.25$ vs. no evidence	1,000	2.167	2.191	+0.024	2/0/998
Retrieved-as-proposal + MiniCheck, $t = 0.25$ vs. raw retrieved	1,000	6.654	2.191	-4.463	60/213/727

Table 11: Dream-v0-Instruct-7B scope check on full-dev ALCE-QAMPARI. Retrieved-evidence decoding is strongly positive against no-evidence decoding, but applying the unchanged MiniCheck admission script to Dream's retrieved output as a proposal source is nearly inert and far below raw retrieved Dream output. This is appendix scope evidence, not a proposal-admission transfer result.

Admission rule	Evidence	Threshold	Δ F1	W/L/T	Out. Chg.	Admitted slots
MiniCheck-Flan	real	0.10	+0.355	13/0/987	39	88
MiniCheck-Flan	real	0.15	+0.364	13/0/987	32	71
MiniCheck-Flan	real	0.25	+0.356	12/0/988	28	62
MiniCheck-Flan	real	0.35	+0.313	11/0/989	25	56
MiniCheck-Flan	shuffled	0.10	0.000	0/0/1000	0	0
MiniCheck-Flan	shuffled	0.25	0.000	0/0/1000	0	0
WeCheck	real	0.25	+0.421	14/0/986	40	96
WeCheck	real	0.50	+0.393	14/0/986	32	73
WeCheck	real	0.60	+0.360	13/0/987	27	61
WeCheck	real	0.70	+0.366	13/0/987	27	57
WeCheck	shuffled	0.25	0.000	0/0/1000	8	22
WeCheck	shuffled	0.60	0.000	0/0/1000	0	0
AlignScore-large	real	0.25	+0.424	14/0/986	49	117
AlignScore-large	shuffled	0.25	+0.453	14/0/986	63	205
SummaCConv	real	0.05	+0.542	18/0/982	83	282
SummaCConv	shuffled	0.05	+0.542	18/0/982	83	282

Table 12: Threshold and shuffled-evidence stability. Out. Chg. counts true output changes; admitted slots counts slots accepted into the output. Local admission rules retain positive real-evidence gains while their shuffled-evidence admissions collapse at the selected operating points. Holistic factuality scorers can retain high gains and high admitted-slot counts under shuffled evidence.

Operating point	Δ F1	95% CI	W/L/T	Sign-test p
Proposal-only	+0.539	[+0.273,+0.865]	18/0/982	7.63×10^{-6}
PAD (MiniCheck-Flan)	+0.356	[+0.147,+0.611]	12/0/988	4.88×10^{-4}
PAD (WeCheck)	+0.360	[+0.161,+0.609]	13/0/987	2.44×10^{-4}
PAD (MiniCheck-RoBERTa)	+0.299	[+0.107,+0.550]	10/0/990	1.95×10^{-3}
PAD (MiniCheck-DeBERTa)	+0.255	[+0.087,+0.452]	9/0/991	3.91×10^{-3}
PAD (deterministic)	+0.292	[+0.107,+0.508]	9/0/991	3.91×10^{-3}

Table 13: Full-dev QAMPARI public-metric statistics for the main positive operating points.

DeBERTa gives $\Delta F_1 = +0.255$ with 95% CI $[+0.087, +0.452]$ and 9/0/991 W/L/T (sign-test p=3.91×10^{-3}). With four independently trained NLI verifiers (a Flan-T5-large encoder, a RoBERTa-large encoder, a DeBERTa-v3-large encoder, and the WeCheck weakly supervised verifier) all yielding small, positive, statistically non-random gains with zero losses, the operating point is not tied to a single verifier checkpoint.

C.12 Consensus Admission Is a Conservative Reference

We also tested whether combining admission rules by consensus improves the operating point. Requiring both MiniCheck-Flan and WeCheck to admit a slot yields +0.319 F1 with 11/0/989 W/L/T. A two-of-three consensus over MiniCheck-Flan, WeCheck, and the deterministic reference yields +0.331 F1 with 11/0/989 W/L/T. Both variants are positive and loss-free, but they are weaker than the single local-admission operating points. We therefore report consensus as a conservative reference rather than making it the method identity.

C.13 Proposer Ablation

The canonical Proposal-Admission Decoding operating point uses a selective-repair SPREAD-style proposer. We ask whether trajectory-level contrastive decoders can substitute as proposers under the same MiniCheck-Flan admission rule at threshold 0.25. Table 14 compares three proposers paired against the same retrieved-evidence baseline. A-CFG-style and CAD-D outputs barely deviate from the retrieved baseline at their reported operating points, so MiniCheck admission has almost no novel candidates to evaluate. Both yield essentially flat operating points (ΔF_1 of $+0.020$ and 0.000 with at most 3 admitted slots), in contrast to the SPREAD-style proposer that supplies 62 admitted slots and $+0.356$ ΔF_1. This supports the paper's choice of a selective-repair SPREAD-style proposer: the admission boundary needs a

Proposer	Admitted slots	Out. Chg.	Δ F1	95% CI	W/L/T
SPREAD (selective repair)	62	28	**+0.356**	[+0.147, +0.611]	12/0/988
A-CFG-style (w=0.5, ρ=0.7)	3	3	+0.020	[+0.000, +0.060]	1/0/999
CAD-D (α=0.35)	2	2	+0.000	[+0.000, +0.000]	0/0/1000

Table 14: Proposer ablation under fixed MiniCheck-Flan admission at threshold 0.25. Each row is paired against the same LLaDA-8B retrieved-evidence baseline on QAMPARI full-dev (n=1,000). Green cell marks the canonical operating point; red cells mark proposers whose admission is essentially inert.

	MiniCheck-Flan (τ=0.25)		WeCheck (τ=0.60)	
Condition	Accept rate	AUROC	Accept rate	AUROC
Original	0.600	–	0.688	–
Evidence swap	0.000	0.947	0.000	0.950
Query swap	0.000	0.961	0.000	0.982
Slot swap	0.312	0.686	0.350	0.732
Hard negative	0.562	0.540	0.487	0.613

Table 15: Slot-binding counterfactual results at the canonical thresholds. Accept rate is the fraction of perturbed cases the verifier still admits; AUROC is the discrimination between original and perturbed scores. Green cells mark conditions where the verifier correctly rejects all perturbations; red cells mark conditions where rejection is incomplete. Both verifiers track evidence- and query-binding but are weaker on slot surface and nearby hard-negative substitutions.

permissive proposer that produces candidate slots the retrieved baseline does not already contain.

C.14 Slot-Binding Counterfactuals

We test whether local admission uses the retrieved evidence and query rather than only slot plausibility. For accepted candidate slots, we construct four counterfactual conditions: evidence swaps, query swaps, slot swaps, and nearby hard negatives. Table 15 reports accept rates and AUROCs for both verifiers. MiniCheck-Flan rejects all evidence-swap and query-swap cases at threshold 0.25; WeCheck does the same at threshold 0.60. The corresponding AUROCs are 0.947 and 0.961 for MiniCheck-Flan, and 0.950 and 0.982 for WeCheck. Both verifiers remain weaker on slot swaps and hard negatives: MiniCheck-Flan accepts 31.2% of slot swaps and 56.2% of hard negatives, while WeCheck accepts 35.0% and 48.7%. These failures motivate reporting slot-binding counterfactuals in addition to shuffled-evidence tests.

D Qualitative Cases and Slot-Binding Counterfactuals

This appendix provides qualitative and counterfactual analyses of the slot-level admission decisions summarized in Section 5.5. These analyses are not used as the primary evidence for the public-metric gains; instead, they localize the remaining admission failures after candidate generation and final inclusion have been separated.

Changed cases. Changed cases fall into three recurring patterns: valid self-answers that require relation-sensitive admission, plausible entities whose local frames do not license the queried relation, and homonym or inverted-relation candidates surfaced by proposal-only decoding. These cases show that admission should test query-slot relation binding rather than surface plausibility. The examples below provide full pipeline traces for representative cases.

Slot-binding counterfactuals. Counterfactual binding tests hold the proposal set fixed while replacing one binding dimension at a time: the retrieved evidence, the query, the slot surface, or a nearby hard negative. At their operating thresholds, MiniCheck-Flan and WeCheck reject all evidence-swap and query-swap cases (AUROC $>$ 0.94 and $>$ 0.96), but accept 31–35% of slot swaps and 49–56% of hard negatives. These failures identify fine-grained near-miss discrimination as the remaining admission bottleneck: once proposal and admission are separated, errors can be attributed to specific slot-level decisions rather than to the whole denoising trajectory. Aggregate counterfactual results are reported in Appendix C.14.

We work through four representative changed cases from the full-dev QAMPARI run. Each case shows the question, the selective-repair proposer's candidate slots, and the MiniCheck-Flan admission scores at the canonical operating threshold τ=0.25. All numbers are taken from the unmodified run and public ALCE evaluation traces; no values are hand-constructed.

D.1 Case A: Director-Writer Binding

This is the case visualized in Figure 3.

Query. For which movie did Gregory Nava serve dual roles as director and writer?

Baseline output. `insufficient evidence.`

Proposer candidates and admission scores.

- **Bordertown** score 0.795 ✓ admit
- American Justice score 0.232 × reject (below τ)

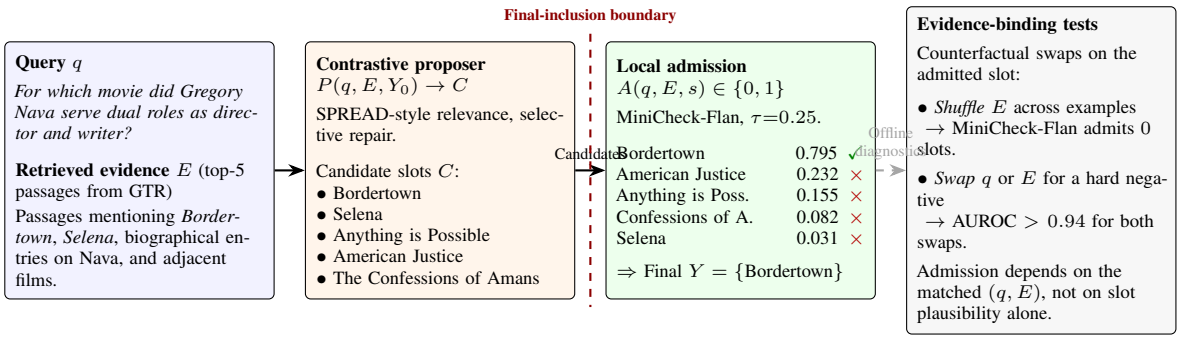

Figure 3: Proposal-admission pipeline on a changed case from the full-dev QAMPARI run. A SPREAD-style proposer supplies five candidate slots from the retrieved evidence; MiniCheck-Flan admission at τ=0.25 accepts only the slot whose local evidence binds the queried director-and-writer relation. All MiniCheck-Flan scores are from the unmodified run.

Question pattern	Candidate	Local evidence observation	Admission behavior
Movie directed by the person who wrote a Dan Gilroy screenplay	Dan Gilroy	The retrieved frame states that Gilroy wrote and directed *Nightcrawler*. The candidate repeats the query entity, but the same local frame binds both roles.	Admit
Colleges attended by employees of Girton College	University of Cambridge; Girton College	The retrieved frame states that Girton is a Cambridge constituent college. It does not state where Girton employees studied.	Reject or leave unchanged
Education of Second Circuit judges	Harvard Law	The retrieved frame describes the court and lists judges. The school is plausible, but the frame does not bind the queried education relation to the listed people.	Reject or leave unchanged
TV show directed by someone Tom Kauffman wrote for	Lloyd Kaufman; George S. Kaufman	The frame mentions different Kaufman/Kauffman names, but none binds Tom Kauffman as the screenwriter to the proposed directors.	Reject
Representative of the Roman Republic	Pope Pius IX	The frame says Pope Pius IX fled the Roman Republic and names other leaders. The proposed slot reverses the relation.	Reject

Table 16: Representative changed cases from the full-dev QAMPARI run. The first case shows a valid self-answer that broad query-entity rejection would remove; the remaining cases show false-positive candidates that proposal-only decoding can supply but slot-level admission should reject.

- Anything is Possible score 0.155 × reject
- The Confessions of Amans score 0.082 × reject
- Selena score 0.031 × reject

The admitted slot is *Bordertown*, where the local evidence frame binds Nava to both director and screenwriter roles. *American Justice*, at score 0.232, sits just below the admission threshold. The retrieved frame contains nearby filmography and writing-collaboration context, but it does not bind this title to both the director and writer roles required by the query. We therefore treat the score as a near-threshold ambiguity, not as evidence of reliable relation calibration. *Selena*, scored 0.031, may be plausible from world knowledge but is not supported by the retrieved passages, so the verifier refuses to admit it. This illustrates the intended boundary: admission is grounded in the local passages, not in the verifier's parametric knowledge.

D.2 Case B: Stoke City Transfer Candidates

Query. Which English football player transferred to Stoke City in the summer of 2013?

Baseline output. `insufficient evidence.`

Proposer candidates and admission scores.

- **Cameron Jerome** score 0.865 ✓ admit
- **Matthew Etherington** score 0.860 ✓ admit
- Jordan Richardson score 0.115 × reject
- Karim Rossi score 0.082 × reject
- Jordan Keane score 0.074 × reject
- Ben Heneghan score 0.063 × reject

- Lucas Dawson score 0.054 × reject
- Kevin Gomez-Nieto score 0.053 × reject
- Stephen Kelly score 0.031 × reject

The proposer is permissive and supplies nine candidate players. Two scored well above the threshold and were admitted; the remaining seven are nearby players mentioned in the retrieved Stoke City passages but not bound to the 2013 transfer relation. One passage explicitly binds Cameron Jerome and Matthew Etherington to the 2013 transfer window, while the remaining names appear in adjacent context; the verifier separates these cases on the same retrieved passages.

D.3 Case C: Multi-Role Self-Answer

This is the Dan Gilroy case referenced in Table 16. The behavior is more nuanced than the figure alone shows.

Query. Who directed the movie that had a screenplay written by Dan Gilroy?

Baseline output. `insufficient evidence.`

Proposer candidates and admission scores.

- **Dan Gilroy** score 0.491 ✓ admit (self-answer)
- **Tony Gilroy** score 0.958 ✓ admit

A broad query-entity rejection rule would remove Dan Gilroy because the candidate repeats a query mention. The local evidence frame, however, binds Gilroy to both the screenwriter and director roles, and the verifier admits the candidate with a moderate score of 0.491. The verifier also admits Tony Gilroy at a higher score because Tony directed *The Bourne Legacy*, for which Dan Gilroy is credited as co-screenwriter, and this dual-name binding is present in the retrieved passages. This shows that local evidence binding can trigger admission even when a shallow query-copy heuristic would reject the candidate.

D.4 Case D: Changed-Zero Tie

Query. Which German offshore wind farm was commissioned after 2010?

Baseline output. `insufficient evidence.`

Method output and admission score.

- **Alpha Ventus Offshore Wind Farm** score 0.371 ✓ admit

Official F1 remains 0 because the added slot does not match the gold answer set, so the paired W/L/T summary records a tie. The output nevertheless changes from abstention to a concrete answer. The local frame mentions Alpha Ventus and temporal facts about German offshore wind development, but it does not establish the requested after-2010 commissioning relation. This is the changed-zero failure mode: W/L/T treats the example as behaviorally neutral even though admission has introduced a weakly supported answer. We therefore track changed-output counts alongside W/L/T in the diagnostic analyses.