Trajectory Self-Feedback in Diffusion Language Models: A Systematic Survey

Yash Shah
Arizona State University
yshah124@asu.edu

Abhijit Chakraborty
Arizona State University
achakr40@asu.edu

Vivek Gupta
Arizona State University
vgupt140@asu.edu

Abstract

Diffusion language models generate text through iterative denoising, allowing intermediate predictions to influence later denoising steps, a recurring practice we call trajectory self-feedback (TSF). TSF is adopted under inconsistent terminology, evaluated on incompatible benchmarks, and studied by disconnected communities, yet no prior review characterises it reproducibly. We present a systematic literature review under an evidence-based methodology: six pre-specified questions, a documented six-library search with snowballing, explicit selection and quality criteria, and a structured synthesis of **24 primary studies** across seven clusters, with a released audit trail. We find: (i) a model-agnostic taxonomy organises the reviewed mechanisms into four operator families: Prediction-Space Concatenation, Confidence-weighted Path Signals, Reward-Reinforced Conditioning, and Correction-Targeted Feedback, which the literature often conflates as *self-conditioning*; (ii) evidence of benefit is real but *fragmented* across incomparable settings; (iii) within our search window, no included study jointly varies the TSF mechanism and measures factual faithfulness on knowledge-intensive tasks; and (iv) a discrete/continuous asymmetry motivates an untested sampling-wall hypothesis. We conclude with eight open problems, a dependency map, and review-derived diagnostics (Commitment Index, Correction Capacity) and a Standardised Evaluation Protocol (SEP-DLM) to make TSF measurable across quality, correction, trajectory commitment, and faithfulness.

1 Introduction

Language models increasingly observe and revise their own intermediate outputs without an external judge. In autoregressive (AR) models this is a *post-hoc* loop: self-refinement (Madaan et al., 2023) and self-critique (Bai et al., 2022) re-read an output and emit a revision, but left-to-right generation makes early errors propagate irreversibly, so revision must run externally; recent causal work frames this as *asymmetric attractor dynamics*, where early tokens anchor generation onto a consistent but wrong path (Akarlar, 2026). In this paper, *faithfulness* refers to factual accuracy or hallucination behaviour on knowledge-grounded tasks; n-gram overlap, perplexity, and general fluency are treated as quality metrics rather than factual-faithfulness measures. We ask whether and how TSF— conditioning denoising on the model's own prior prediction—has been evaluated as a route to factual faithfulness in diffusion language models.

Diffusion language models (DLMs) generate text by iteratively denoising a complete parallel sequence, so they observe the entire partial output at every reverse step (Austin et al., 2021; Sahoo et al., 2024; Nie et al., 2025). This permits *intrinsic* self-feedback: each step conditions on the model's own previous prediction and may revise any position before generation is final. The opportunity is not uniform: discrete masked DLMs sample a hard token each step, collapsing a distribution to a one-hot choice and compounding errors downstream (Duckworth et al., 2019; Jo et al., 2025; Hersche et al., 2025), a bottleneck we call the *sampling wall* (defined formally in §2.2). Continuous and flow-matching models keep the sequence continuous, preserving distributional uncertainty (Chen et al., 2026; Jo and Hwang, 2025). The implications of this divide for self-feedback, correction, and faithfulness are almost unexplored (Guo and Tan, 2026). Regardless of family, a common practice has emerged: at reverse step t the model receives the noisy state x_t and a function of its own previous prediction $\tilde{x}_{t+1}$. Introduced for images as *self-conditioning* (Chen et al., 2022) and prefigured in AR LMs by belief routing (Suau et al., 2021), it now appears in DLMs under many names: text generation (Mahabadi et al., 2023), post-training adaptation (Cardei et al., 2026), few-step inference

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37189–37208
October 24-29, 2026 ©2026 Association for Computational Linguistics

(Nguyen-Cong et al., 2026), planner-based decoding (Peng et al., 2025b,a), error correction (Schiff et al., 2026; Zhang et al., 2025; Kim et al., 2025), and continuous flow models (Chen et al., 2026). We adopt *trajectory self-feedback* (TSF) as the umbrella term, distinct from conditioning on external signals. This has created three disjoint sub-fields that study TSF without citing one another: *self-conditioning/efficiency* work evaluates perplexity and n-gram quality, rarely hallucination; *correction* work repairs code or synthetic corruptions without connecting to trajectory-level failure; and *hallucination* work studies trajectory failure in *default* DLMs without varying the mechanism (§5).

Why a *systematic* review. Broad surveys of diffusion and non-autoregressive text generation (Li et al., 2023, 2025; Yu et al., 2025) provide horizontal landscape maps, while AR self-correction surveys such as Pan et al. (2024) address a different feedback setting; none provides an auditable TSF-specific denominator across DLM families. We therefore adopt the Evidence-Based Software Engineering systematic-review methodology (Kitchenham et al., 2015, 2007) containing a pre-specified review protocol, repeatable search, explicit criteria, structured extraction, and threats analysis, which turns "three communities do not talk" into a reproducible finding and "no paper varies TSF and measures faithfulness" into a falsifiable claim with a documented denominator. Our contributions are: (1) a **problem-driven SLR** of TSF under Kitchenham et al. (2015) comprising six pre-specified RQs, a six-library search, a PRISMA flow over 24 studies, and a released audit trail (§3); (2) a **model-agnostic taxonomy** of four operator families across discrete, continuous, and flow DLMs (§4); (3) a **seven-dimension gap synthesis** with a per-study matrix (§5); (4) **eight open problems** with a dependency map (§6); and (5) the CI/CCap **diagnostics and SEP-DLM** (§7), together with a released single-backbone reference run that instantiates the protocol end to end as feasibility evidence rather than broad validation.

2 Background

2.1 Diffusion language models

DLMs apply diffusion to text via a forward noising process and a learned reverse denoiser; three families scope this review. **Discrete masked DLMs** make categorical per-step predictions: D3PM (Austin et al., 2021) (structured/absorbing state processes), SEDD (Lou et al., 2023) (score-entropy ratios), MDLM (Sahoo et al., 2024) (single `[MASK]` absorbing state, parallel decoding), and LLaDA (Nie et al., 2025) (LLM scale). For post-corpus context, LLaDA2.1 (Bie et al., 2026) studies faster text diffusion via token editing; it post-dates the corpus refresh and therefore does not alter any operator-family assignment or corpus-level count. **Continuous text diffusion** embeds tokens in $\mathbb{R}^D$, adds Gaussian noise, and denoises before projecting: DiffuSeq (Gong et al., 2022) (seq-to-seq), RDLM (Jo and Hwang, 2025) (Riemannian), and LangFlow (Chen et al., 2026) (flow-matching with a cross-entropy objective over noisy embeddings). **Flow-based LMs** learn a velocity field from prior data: FLM/FMLM (Lee et al., 2026) (flow-matching over one-hot, one-step distillation) and DLM-One (Chen et al., 2025) (one-step distillation of continuous DLMs). At the paradigm level, masked diffusion can outperform AR models in data-constrained regimes through implicit data augmentation over token orderings (Prabhudesai et al., 2025), indicating that DLM advantages are regime-dependent and motivate closer scrutiny of where they hold.

2.2 The sampling wall

At each step, a standard *masked diffusion language model* (MDLM) samples categorically, collapsing the distribution to one hard token; a wrong sample passes collapsed one-hot evidence forward as if correct, compounding error, which we call the *sampling wall*. Architectural responses include latent propagation (Jo et al., 2025), soft-masking via embedding superposition (Hersche et al., 2025), continuous/flow models that defer projection (Chen et al., 2026; Lee et al., 2026), and draft-then-refine remasking (Ma et al., 2026). This bears directly on TSF: discrete MDLMs feed back a collapsed one-hot sample, whereas continuous and flow models carry a full distributional prediction. We state this as a falsifiable conjecture, not an established result:

> **Sampling-wall hypothesis ($\mathbf{H_{wall}}$).** Hard categorical commitment in discrete MDLMs is a primary driver of the observed discrete/continuous asymmetry in TSF effectiveness; softening that commitment should reduce early trajectory commitment and error amplification.

Epistemic status. H_{wall} is an *untested correlational interpretation*: no included study isolates categorical commitment as the causal variable. We use

it only to motivate falsifiable comparisons; downstream observations are described as consistent or inconsistent with the hypothesis, not as evidence that establishes it.

2.3 The self-feedback pattern

Let f_θ be the denoiser. Without self-feedback $\tilde{x}_t = f_\theta(x_t, t, c)$ for external conditioning c (e.g. a source sentence or a prepended AR plan); with trajectory self-feedback,

$$\tilde{x}_t = f_\theta(x_t,\, t,\, c,\, g(\tilde{x}_{t+1}, x_t))\,, \qquad (1)$$

where g is the *self-feedback operator*. The fed-back $\tilde{x}_{t+1}$ is a one-hot sample (discrete), an embedding (continuous), or a velocity prediction (flow), so g is realised differently across families even when the pattern is identical; the choice of g defines the mechanism family (§4). For families where g produces a distribution over token positions that modifies the unmasking schedule rather than entering the denoiser directly (CPS, §4), Eq. (1) denotes a generalised conditioning pattern: g maps $(\tilde{x}_{t+1}, x_t)$ to a signal s_t, which acts on the decoding schedule rather than on f_θ's input; the operator boundary is preserved. TSF is distinct from trajectory self-distillation and external plan conditioning, which fall outside Eq. (1) (boundary references, §8).

3 Review Methodology

We follow the three-phase *evidence-based software engineering* (EBSE) process of Kitchenham et al. (2015, 2007) for *planning* (research questions, protocol), *conducting* (search, selection, quality assessment, extraction, synthesis), and *reporting*, along with PRISMA 2020 (Page et al., 2021) for selection-flow reporting and Wohlin's guidelines (Wohlin, 2014) for snowballing. Because TSF is a mechanism rather than an effect-size question, we conduct a *qualitative* SLR with structured narrative synthesis (justified in §3.7), not a meta-analysis.

3.1 Research questions

The review is organised around six pre-specified questions:

RQ1 What forms of TSF exist across DLM families, and can they be unified under one operator taxonomy?

RQ2 What evidence exists for TSF's effect on quality, efficiency, and error correction?

RQ3 How is TSF evaluated, and are results comparable across studies?

RQ4 Does TSF's effect differ across discrete, continuous, and flow families?

RQ5 Has any study connected TSF design to hallucination or faithfulness on knowledge-intensive tasks?

RQ6 What diagnostics and protocols would make TSF evidence comparable and cumulative?

RQ1 is addressed in §4; RQ2–RQ5 in §5; RQ6 in §7 (Review-Derived Evaluation Recommendations).

3.2 Search strategy

Sources. We searched six digital libraries: arXiv, ACL Anthology, OpenReview, and Semantic Scholar via API/import, and IEEE Xplore and ACM Digital Library by manual search owing to public API restrictions (Appendix B); DBLP and Google Scholar were reserved for snowballing. Database **search window.** January 2021 (D3PM (Austin et al., 2021)) through March 2026. A pre-camera-ready forward-snowballing refresh in April 2026 identified SCMDM (Cardei et al., 2026), which was screened under the same criteria and added to the final corpus. **Search string.** From a PICO decomposition (Population: diffusion language models; Intervention: self-feedback; Outcome: quality, correction, faithfulness), the canonical Boolean string was:

```
(diffusion) AND ("language
model" OR "text generation"
OR "sequence generation")
AND (self-condition* OR
self-feedback OR self-correct*
OR remask* OR "planner" OR
"trajectory" OR "intermediate
prediction" OR "denoising
feedback")
```

Library-specific adaptations are in Appendix A; the string ran in a high-recall configuration, narrowed downstream against the TSF-specific criteria (§3.3).

Snowballing. One backward/forward iteration (Wohlin, 2014) over studies passing screening. **Seed-recall check.** As a floor-level retrieval audit, we tested the complete six-library-plus-snowballing pipeline against five foundational papers that a viable search should recover: D3PM (Austin et al., 2021), Analog Bits (Chen et al., 2022), SEDD (Lou et al., 2023), MDLM (Sahoo et al., 2024), and LLaDA (Nie et al., 2025). All five were recovered (5/5). This small known-item set is a sanity check, not a comprehensive estimate of

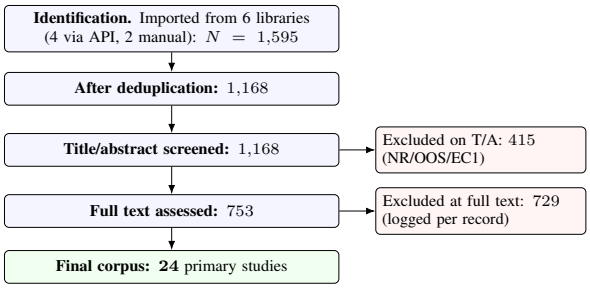


Figure 1: PRISMA-style selection flow. The expanded search is high-recall, so a large pool is narrowed to 24 primary studies. Title/abstract exclusions are coded by reason (Table 3: no DLM/trajectory relevance, out-of-scope, image/audio/video); the 729 full-text exclusions are recorded individually in `screening_sheet.csv`. The 753-record full-text pool includes candidates from one snowballing iteration and the pre-camera-ready refresh.

recall, and is reported alongside the corpus's strong preprint and recency skew.

3.3 Inclusion Criteria (IC)

A study was **included** if it (IC1) concerned a diffusion- or flow-based *text/language* model; (IC2) introduced, used, analysed, or explicitly declined a self-feedback mechanism in the sense of Eq. (1) (or analysed the denoising trajectory bearing on it); and (IC3) was archival or arXiv-deposited. It was **excluded** if (EC1) image/audio/video diffusion only; (EC2) autoregressive with no DLM component (kept as related work); (EC3) a non-archival abstract, tutorial, or blog post; (EC4) a superseded duplicate; or (EC5) a survey (related work, §8).

3.4 Selection process

Figure 1 reports the PRISMA flow. The expanded high-recall search yielded **1,595** imported records; after de-duplication, **1,168** remained. Title/abstract screening against IC1 to IC3 excluded **415**: 290 had no DLM/trajectory-feedback relevance, 67 were otherwise out of scope, and 58 concerned image/audio/video or non-language diffusion (Table 3), leaving **753** records for full-text assessment. This pool includes candidates surfaced by the documented backward/forward snowballing and pre-camera-ready refresh. Full-text assessment excluded **729** records, each logged individually with a recorded reason, yielding a final corpus of **24 primary studies**.[1]

[1]The complete per-record log, per-library hit counts, and exclusion codes are released in the replication repository and summarised in Appendices B to F.

3.5 Quality assessment

Each study was scored on $\{0, 0.5, 1\}$ against eight criteria (Table 4) probing general rigour (QA1, QA4, QA8) and TSF-specific reporting (QA2, QA3, QA5 to QA7), using a written rubric with one calibration example per level. Two reviewers independently scored all 24 studies; quadratic-weighted agreement averaged κ_w=0.89 (range 0.56–1.00 across criteria), lowest for QA7 and QA3, reconciled to consensus (Appendix B). Scores expose reporting gaps (notably QA5/QA6) but do not exclude studies. On the QA5 boundary: PPL/MAUVE/BLEU alone score 0; hallucination rate, FactScore (Min et al., 2023), or trajectory-level AUROC on a knowledge task scores 1. QA6 scores 1 only for measured *repair* of injected or natural errors, not mere quality gains. Per-study scores: Appendix F.

3.6 Data extraction

We extracted seven aspects per study: cluster, self-feedback form, measurement criteria, central hidden assumption, main blind spot, unresolved issues, and QA score. One author extracted all studies, and a second author independently re-coded the most debatable field, operator family. Agreement was κ=0.71 (raw 83%) before the decision rules of Appendix C were formalised, the disagreements concentrating on the PSC-vs.-subtype and CPS-vs.-PSC boundaries, and κ=0.89 after re-coding.

3.7 Synthesis approach

We use *structured narrative synthesis*, recommended for heterogeneous studies with incommensurable outcomes (Kitchenham et al., 2015). Meta-analysis is *infeasible*, which is itself a finding (RQ3): the corpus spans PPL (Ren et al., 2023), MAUVE (Pillutla et al., 2021), BLEU, `pass@k`, foldability, AUROC, and ad hoc proxies with no shared task or protocol. We synthesise by thematic cluster (§5) and research question, culminating in the open-problem formalisation (§6).

4 RQ1: A Taxonomy of TSF Mechanisms

We define a *self-feedback operator* as any map $g : (\tilde{x}_{t+1}, x_t) \mapsto s_t$ producing the conditioning signal s_t in Eq. (1). Its input $\tilde{x}_{t+1}$ differs across families (one-hot token, embedding, or velocity prediction). The four families below cover the mechanisms observed in the *reviewed corpus under our explicit decision rules*; they are not a claim

that all future TSF designs must belong to exactly one of four closed-world categories (Table 1). The same four families can also be recovered top-down by crossing two design axes—what the feedback acts on, and how the fed-back prediction is weighted—which additionally exposes empty, open cells of the design space (Appendix C). We scope the taxonomy to text/language DLMs (IC1); self-conditioning in *image* diffusion (Xiang et al., 2025) and *belief routing* in AR LMs (Suau et al., 2021) share the loop on different signal spaces and are treated as boundary references (EC1/EC2, §8).

Family 1: Prediction-Space Concatenation (PSC). *Analog Bits (Chen et al., 2022), TESS (Mahabadi et al., 2023), SCMDM (Cardei et al., 2026), LangFlow (Chen et al., 2026), LDDM (Jo et al., 2025).* PSC feeds the full previous prediction back as denoiser input, $s_t = \mathrm{concat}(\tilde{x}_{t+1}, x_t)$. It improves PPL and quality in continuous DLMs (Chen et al., 2026), but in discrete MDLMs, it improves quality metrics while slightly degrading PPL— a cross-family asymmetry that $\mathrm{H}_{\mathrm{wall}}$ could explain but that no included study causally tests, and can hurt few-step (Nguyen-Cong et al., 2026). Yet SCMDM's *post-training* two-pass PSC *improves* generative PPL ($43 \rightarrow 24$, OpenWebText) (Gokaslan and Cohen, 2019), so outcomes depend on the training regime (OP2, OP3), not the family alone. *Sub-type (soft-masking)* (Hersche et al., 2025): a confidence-weighted superposition of mask and top-k embeddings lets partial hypotheses propagate without collapse.

Family 2: Confidence-weighted Path Signals (CPS). *P2 (Peng et al., 2025b), PAPL (Peng et al., 2025a).* CPS turns confidence into a distribution over token *positions*, $s_t = \mathrm{softmax}_\tau(p_\theta(\tilde{x}_{t+1}))$, where $\mathrm{softmax}_\tau$ is a temperature-τ softmax over token positions (low τ sharpens routing toward confident positions), routing the unmasking planner rather than entering the denoiser and decoupling *what* is predicted from *where* it updates. This distinguishes CPS from PSC within the Eq. (1) schema: the mapping $g \rightarrow s_t$ is the same; what changes is whether s_t enters f_θ as input (PSC) or modifies the step-selection distribution (CPS). It improves protein foldability and code `pass@k` (Peng et al., 2025b) and text MAUVE/GenPPL (Peng et al., 2025a). Calling CPS "self-conditioning" wrongly conflates it with PSC.

Family 3: Reward-Reinforced Conditioning (RRC). *TREC (Liu et al., 2024).* RRC currently has **one member in the reviewed corpus**. We retain it as a family because reward-scaling is a distinct axis of the top-down design space, not because of instance count; the family would grow if reward- or RL-from-feedback mechanisms are ported into additional DLM training or decoding settings. RRC weights feedback by an advantage estimate, $s_t = A(\tilde{x}_{t+1}) \cdot h(\tilde{x}_{t+1})$, upweighting informative signals; it prevents self-conditioning degradation in training and eases the train–inference mismatch (OP3). Unlike PSC it passes a reward-scaled representation, which could, in principle, dampen error amplification in discrete models; this remains untested.

Family 4: Correction-Targeted Feedback (CTF). *ProSeCo (Schiff et al., 2026), CDLM (Zhang et al., 2025), PRISM (Kim et al., 2025).* CTF pairs $\tilde{x}_{t+1}$ with an explicit error signal flagging likely-wrong positions, $s_t = (\tilde{x}_{t+1}, e_{t+1})$, learned via mixture/joint losses or fine-tuning, to break the wrong-prediction reinforcement loop. It improves code-revision correction (Zhang et al., 2025), GSM8K/code accuracy under refinement (Cobbe et al., 2021; Schiff et al., 2026), and gives provable guarantees under specific fine-tuning (Kim et al., 2025).

Decision procedure and boundaries. Family assignment follows four ordered questions (full rules, Appendix C): does g feed back the full prediction as denoiser input (PSC); convert it into a distribution over token positions that routes the planner (CPS); scale it by a reward/advantage estimate (RRC); or pair it with an explicit error signal (CTF). The hardest boundary in practice is CPS-vs.-PSC: P2 (Peng et al., 2025b) reads as "self-conditioning," but its confidence signal selects *which positions* unmask rather than entering the denoiser as input, so it is CPS, not PSC (rule D2). Trajectory self-distillation (T3D) and external plan conditioning fall *outside* Eq. (1) and are treated as boundary references (§8). Hybrid operators that combine channels (e.g. RemeDi (Huang et al., 2025)) are assigned by the primary channel on which the training loss is computed, with auxiliary channels recorded as annotations (Appendix C).

Family	Instantiations	Signal space	Used for	Known benefit	Known failure
PSC	Analog Bits, TESS, SCMDM, LangFlow, LDDM	Full prediction (one-hot or embedding)	Quality	Helps continuous DLMs; discrete PPL regime-dependent	Wall amplification; degrades few-step (Nguyen-Cong et al., 2026)
PSC subtype: SM	Soft-Masked DLM	Continuous embedding superposition	Sampling-wall mitigation	Removes hard-commitment error in discrete	Studied only in discrete; no faithfulness eval (Hersche et al., 2025)
CPS	P2, PAPL	Confidence dist. over positions	Path routing	Improves planning/alignment	No faithfulness/correction eval (Peng et al., 2025b,a)
RRC	TREC	Reward-weighted signal	Conditioning stability	Prevents train. degradation	Seq2seq only; correction unstudied (Liu et al., 2024)
CTF	ProSeCo, CDLM, PRISM	Prediction + error signal	Error correction	Improves correction on code/math	Code/synthetic only; no factual QA (Zhang et al., 2025; Schiff et al., 2026)

Table 1: Taxonomy of TSF operator families (RQ1). All four families improve their target metric; *none* has been evaluated for faithfulness on knowledge-intensive tasks. "Self-conditioning" denotes PSC-type mechanisms; CPS, RRC, and CTF are distinct operator families and should not be labelled self-conditioning. Per-study conditions, references, and metrics are in Table 8 and §5; we do not pool effect sizes because the setups are incommensurable (RQ3), so benefit/failure entries are directional rather than meta-analytic estimates.

5 Findings: Cross-Cluster Synthesis

We synthesise by the seven thematic clusters that emerged during extraction (gap matrix, Appendix Table 9); RQ2–RQ5 are answered cumulatively. The cross-cluster tensions the clusters expose when read together—quality-optimising work (C2/C3) measuring something other than factual failure (C5), and correction work (C4) tested almost only on discrete code/math—are discussed in Appendix F.

C1: Foundational DLMs (no TSF). Foundational models (Austin et al., 2021; Gong et al., 2022; Lou et al., 2023; Sahoo et al., 2024; Nie et al., 2025) largely omit self-feedback at training (it complicates the ELBO); MDLM declines self-conditioning and LLaDA only remasks at inference. Their shared assumption, i.e. *independent-step optimality*, is refuted by PAPL (Peng et al., 2025a), which shows the MDLM ELBO is invalid under planned inference.

C2: Self-conditioning & efficiency. These studies (Chen et al., 2022; Mahabadi et al., 2023; Liu et al., 2024; Nguyen-Cong et al., 2026; Hersche et al., 2025; Cardei et al., 2026) define self-conditioning differently and report incomparable metrics; *none* measures hallucination or correction (RQ5). LangFlow (Chen et al., 2026) shows PSC helps continuous DLMs on PPL and quality but discrete ones only on quality. This pattern is consistent with H_{wall} but does not identify categorical commitment as its cause (RQ4). No study evaluates all four families under matched conditions (RQ2, RQ3).

C3: Planner-aware decoding. P2 (Peng et al., 2025b) and PAPL (Peng et al., 2025a) treat unmasking order as a first-class object and show standard training is misspecified for planned inference, but neither asks whether planned trajectories are more faithful nor frames its confidence routing as CPS feedback (RQ1). A training-free *external* counterpoint, plan conditioning (Sauver, 2026), prepends an AR plan for large reasoning gains (GSM8K $75.6 \rightarrow 87.2$, LLaDA-8B; same backbone and evaluation setting, so the number is best read as a within-backbone plan-conditioning gain rather than a general DLM-vs.-AR comparison) but conditions on another model rather than the DLM's own trajectory, so it bounds what TSF must beat while lying outside the definition (§8).

C4: Correction & revision. These studies (Zhang et al., 2025; Schiff et al., 2026; Kim et al., 2025) show correction is *trainable* but share three limits: non-overlapping benchmarks (RQ3), synthetic rather than naturalistic corruption (RQ5), and no link to trajectory dynamics; all study discrete MDLMs only (RQ4).

C5: Hallucination & faithfulness. TraceDet (Chang et al., 2025), DynHD (Qian et al., 2026), and Lost in Diffusion (Guo and Tan, 2026) show DLM trajectories carry exploitable structure and DLM-specific failures (premature termination, context intrusion) absent from AR LMs, but *none* varies the self-feedback mechanism, so they cannot say whether any TSF family helps (RQ5). Post-corpus work has begun to move from trajectory diagnosis toward intervention: OSCAR (Shah et al., 2026a) uses disagreement across parallel denoising chains to localise uncertainty and trigger targeted remasking for hallucination mitigation. Because its cross-chain signal and correction stage do not instantiate the per-step TSF operator in Eq. (1), we treat it as adjacent trajectory-control work rather than part of the systematic corpus.

C6: Continuous & flow LMs. LangFlow (Chen et al., 2026), RDLM (Jo and Hwang, 2025), FLM/FMLM (Lee et al., 2026), and DLM-One (Chen et al., 2025) are assessed on quality and speed, not correction, commitment, or hallucination (RQ4, RQ5). Their continuous/flow state representations avoid the hard categorical commitment posited by H_{wall}, but whether this causes different TSF behaviour is untested.

C7: Sampling-wall mitigation. LDDM (Jo et al., 2025) and Soft-Masked DLM (Hersche et al., 2025) target the wall via latent propagation and soft-masking, reporting iso-compute quality gains but measuring neither correction nor faithfulness, leaving the central hypothesis untested (RQ4, RQ5).

The cross-cutting finding (RQ5). Aggregating QA5 yields the central, falsifiable result: *within our search window (Jan 2021–Mar 2026), no included study varies the self-feedback mechanism and measures faithfulness on a knowledge-intensive task.* Of the 24 studies, **0** both vary TSF and report a faithfulness metric (FactScore, hallucination rate, or factual QA (Thorne et al., 2018); see §1 for the full definition); **3** report faithfulness without varying TSF (the C5 studies); **3** report correction without faithfulness (the CTF studies; Table 8). The released matrix makes the strict claim auditable: one counterexample within the stated definition would falsify it. The denominator is auditable by sub-field and venue (Appendix F).

Faithfulness-boundary sensitivity. The intersection is definition-dependent: **0/24** under the *Strict* factual definition used for RQ5, **1/24** if functional correctness (`pass@k`, foldability) is admitted, and **8/24** if PPL-family quality proxies are admitted. The zero-count headline is therefore specifically a factual-faithfulness result and should not be conflated with the broader readings; the domain-separated breakdown is in Appendix F (Table 10).

6 Eight Open Problems

The synthesis crystallises into eight interlocking open problems.

OP1 (Definition). No shared definition; four mechanisms collapse under "self-conditioning," making comparison unreliable.

OP2 (Theory). No predictive theory of when TSF helps: PSC helps continuous but only partly discrete (Chen et al., 2026), fails few-step (Nguyen-Cong et al., 2026), and can degrade in training (Liu et al., 2024).

OP3 (Train–inference mismatch). TREC, FastDiSS, P2, and PAPL each address a version without recognising it as shared; a unified measure is needed.

OP4 (Correction capacity). CDLM, ProSeCo, and PRISM use incompatible setups; a shared metric spanning code, text, and factual QA is needed.

OP5 (Trajectory commitment). AR attractor dynamics (Akarlar, 2026) have no DLM analogue; a commitment measure must be defined and validated.

OP6 (Faithfulness). TSF and faithfulness are entirely decoupled (RQ5), the most consequential gap.

OP7 (Continuous/flow coverage). Correction/hallucination work is almost entirely discrete; generalisation to continuous/flow LMs is untested. H_{wall} provides one falsifiable prediction for such comparisons but is not itself established evidence.

OP8 (Standardisation). Incommensurable metrics (PPL, MAUVE, BLEU, `pass@k`, foldability, AUROC) make cross-paper claims impossible without a standard protocol.

Dependency map. The problems are ordered (Figure 2): OP1 precedes OP2; OP2/OP3 set the design space for OP4/OP5; OP4/OP5 are the diagnostics for OP6; OP7 requires OP4/OP5 resolved for discrete DLMs first; and OP8 supplies measurement infrastructure throughout — roadmap OP1 → OP8 → OP2/OP3 → OP4/OP5 → OP6 → OP7.

7 Review-Derived Evaluation Recommendations

To address OP4–OP6 and OP8 we propose two diagnostics and a protocol as review-derived measurement recommendations, not validated standards (OP1's shared vocabulary is the taxonomy of §4). A released LLaDA-3B/PopQA reference run (§7.4; Appendix D) demonstrates that the full protocol can be instantiated end to end on one backbone; it establishes feasibility, not general validity.

7.1 Commitment Index (OP5)

Definition 1 (Commitment Index)**.** *Fix a denoiser f_θ and a* reference *run that generates $x^\star$ with self-feedback active throughout (SC_{on}). The Commitment Index at step t_0 is the expected divergence of the final output when,* at step t_0 only*, the self-feedback signal is ablated and the state perturbed:*

$$\mathrm{CI}(t_0) = \mathbb{E}_\delta\big[\, D_{\mathrm{edit}}\big(y_{ref},\, y(x_{t_0}+\delta,\, \mathrm{SC}_{off}@t_0)\big)\big]\,, \tag{2}$$

where $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and y_{ref} is the reference decode; the counterfactual ablates only s_{t_0} and perturbs x_{t_0} before $\mathrm{SC}_{\mathrm{on}}$ resumes. High CI means strong commitment; the curve $\mathrm{CI}(t)$ tests whether discrete models commit earlier than continuous ones (OP7). Perturbation scale, D_{edit}, and channel-only ablation are in Appendix E.

7.2 Correction Capacity (OP4)

Definition 2 (Correction Capacity)**.** *Given evaluation examples with ground truth $x^\star$, each corrupted to $x_{corrupt}$ at error rate ϵ, the Correction Capacity at budget K is*

$$\mathrm{CCap}(\epsilon, K) = \mathbb{E}\left[\frac{\Delta_{correct} - \Delta_{regress}}{\epsilon \cdot |x^\star|}\right], \tag{3}$$

where $\Delta_{\mathrm{correct}}, \Delta_{\mathrm{regress}}$ count repaired and newly-broken positions within K steps. Positive CCap denotes net repair. Under this normalisation the upper value is bounded by complete repair of the injected corruption, but the lower tail is not strictly bounded at -1: refinement can break positions that were originally correct. We therefore report lexical and semantic corruption separately over ϵ and K; full protocol is in Appendix E.

7.3 SEP-DLM (OP8)

The Standardised Evaluation Protocol for DLMs specifies four axes any study claiming a TSF effect should report (Table 11): generation quality, correction capacity, trajectory commitment, and faithfulness, so a gain on one axis can be weighed against a cost on another, enabling the cross-paper comparison RQ3 shows is impossible today. We also recommend *iso-compute* comparison: two-pass families (PSC, CTF) need two forward passes per step and must be matched on total FLOPs, not optimisation steps (Hersche et al., 2025); the full compute-accounting rules, and a reporting checklist, are given in Appendix E, together with the rationale tying each axis to a failure mode the synthesis surfaced.

7.4 Reference feasibility run

In response to the review request to instantiate the proposed instruments, we ran SEP-DLM on LLaDA-3B (Nie et al., 2025) over 400 PopQA examples (Mallen et al., 2023). The remasking decoder and unmasking schedule are held fixed while only feedback strength is varied. At the endpoint comparison, TSF-on vs. TSF-off gives generative PPL 164.9 vs. 610.3, factual accuracy 0.220 vs. 0.098, token-F1 0.127 vs. 0.041, and hallucination rate 0.780 vs. 0.903. Accuracy rises and hallucination falls monotonically across the released strength sweep (Appendix D). The same run instantiates CI with a hard-flip intervention (AUC ≈ 0.12) and CCap with semantic (≈ 0.13) and lexical (≈ 0) corruption. Code and per-axis logs are released alongside the audit trail. This is a *feasibility demonstration*: one backbone, one dataset, and $N{=}400$ do not validate CI, CCap, or SEP-DLM generally. Bootstrap confidence intervals were still pending in the rebuttal release; we therefore make no significance claim from these point estimates or from monotonicity alone.

7.5 A research agenda

The reference run closes the narrow *can SEP-DLM be executed end to end?* question for one mechanism, but not the family-comparison agenda. The highest-priority remaining study (closing OP6) is to *simultaneously* vary TSF family (PSC/CPS/RRC/CTF/none) and measure factual faithfulness on a knowledge-intensive task under SEP-DLM, on one backbone across $T \in \{256, 32, 8\}$. Three follow: cross-backbone validation of CI (including comparison to AR bifurcation/activation-patching (Akarlar, 2026)); a cross-domain CCap benchmark for controlled CDLM/ProSeCo/PRISM comparison (OP4); and a discrete-vs.-continuous-vs.-flow test of $\mathrm{H}_{\mathrm{wall}}$

(OP7). Until these run, family-general faithfulness and diagnostic-validity claims remain open.

8 Related Surveys

Table 12 (Appendix G) compares this survey against five adjacent surveys with explicit scope decisions. Li et al. (2023), Li et al. (2025), and Yu et al. (2025) map broad diffusion/non-autoregressive text or DLM landscapes without systematically isolating trajectory feedback, and Pan et al. (2024) concerns AR/post-hoc correction; ours is vertical by comparison: one mechanism examined across all three DLM families with an auditable denominator. The remaining adjacent literature is cited as boundary references fixing the scope of intrinsic, per-step TSF: AR self-feedback (Madaan et al., 2023; Bai et al., 2022; Zelikman et al., 2022) and AR trajectory commitment (Akarlar, 2026), which supplies the diagnostic template adapted in §7; post-hoc AR self-judging (Lu et al., 2024; Wu et al., 2026); and trajectory control in image diffusion (Kim et al., 2023; Wu et al., 2025). Two adjacent DLM works fall outside Eq. (1)—T3D (Zhang et al., 2026) uses rollouts as a distillation teacher rather than a per-step operator, and plan conditioning (Sauver, 2026) prepends an external AR plan as context c—as does Multi-Path Discrete Consistency (Amin et al., 2026), whose "CDLM" is distinct from the Corrective DLM of Zhang et al. (2025).

9 Conclusion

Reviewing 24 studies under Kitchenham et al. (2015), we identified TSF in seven clusters and three non-communicating communities, with a strict factual-faithfulness gap: *no included primary study jointly varies self-feedback and measures factual faithfulness on a knowledge-intensive task*. The four families cover the reviewed corpus under explicit rules, not all future TSF designs. Our taxonomy, eight open problems, diagnostics (CI, CCap), and SEP-DLM aim to make evidence cumulative and comparable; the released pilot shows the protocol is executable, leaving broad validation and family-level comparisons open.

Limitations

Our conclusions are bounded by the database-search window (January 2021 to March 2026), the inclusion criteria (archival text-DLM papers with a per-step self-feedback component), reliance on recent preprints (22 of 24 studies are arXiv-indexed), and the interpretive nature of operator-family coding. These affect coverage and generalisability rather than the logical structure of the taxonomy or the falsifiability of the gap claim, which is one of *absence of evidence* and is refuted by a single counterexample checked against the released denominator (Appendices B to F). Methodological limits are detailed by category in Appendix H. Three further limits bound our conclusions. First, the corpus is dominated by English, code, and synthetic tasks, so the faithfulness gap may understate failures on languages and domains absent from the included benchmarks; SEP-DLM should therefore specify multilingual, diverse-knowledge benchmarks (App. E). Second, several continuous/flow checkpoints lack public implementations, so the agenda of §7.5 is easier to run on discrete MDLMs—risking reinforcement of the very asymmetry we criticise—and presumes compute not all groups have. Third, promoting TSF as a reliability intervention before faithfulness is measured is risky: a mechanism that improves fluency while leaving factual error unchanged would look like progress under today's metrics and could amplify confident hallucinations, which is why SEP-DLM mandates a faithfulness axis alongside quality and correction. The pilot is one feasibility measurement, not evidence that SEP-DLM or any TSF family reliably improves factuality; its point estimates lack bootstrap confidence intervals and need replication across backbones, tasks, and families.

Ethical considerations

All experiments used publicly available datasets with appropriate licensing and data privacy. The released review artifacts (search log, screening sheet, per-rater quality sheets, extraction matrix, and the CI/CCap reference implementation) are distributed under CC BY 4.0; third-party scholarly artifacts are used in accordance with their platform terms. AI assistance was used for writing clarity only; the authors reviewed all such edits and remain accountable for the final content and conclusions.

References

Gokturk Aytug Akarlar. 2026. Hallucination as trajectory commitment: Causal evidence for asymmetric attractor dynamics in transformer generation. *ArXiv*, abs/2604.15400.

Hasan Amin, Yuan Gao, Yaser Souri, Subhojit Som, Ming Yin, Rajiv Khanna, and Xia Song. 2026. Consistent diffusion language models. In *Forty-third International Conference on Machine Learning*.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *ArXiv*, abs/2107.03006.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback.

Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fu-Chen Chen, Kun Chen, Lun Du, Daozhuo Feng, Haibo Feng, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jian Guan, Kaiyuan Guan, Hongliang He, Zenan Huang, Juyong Jiang, Zhonghui Jiang, Zhenzhong Lan, and 31 others. 2026. Llada2.1: Speeding up text diffusion via token editing. *ArXiv*, abs/2602.08676.

Michael Cardei, Huu Binh Ta, and Ferdinando Fioretto. 2026. Simple self-conditioning adaptation for masked diffusion models. *ArXiv*, abs/2604.26985.

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip H. S. Torr, and Jindong Gu. 2025. Tracedet: Hallucination detection from the decoding trace of diffusion large language models. *ArXiv*, abs/2510.01274.

Tianqi Chen, Shujian Zhang, and Mi Zhou. 2025. Dlm-one: Diffusion language models for one-step sequence generation. *ArXiv*, abs/2506.00290.

Ting Chen, Ruixiang Zhang, and Geoffrey E. Hinton. 2022. Analog bits: Generating discrete data using diffusion models with self-conditioning. *ArXiv*, abs/2208.04202.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. 2026. Langflow: Continuous diffusion rivals discrete in language modeling. *ArXiv*, abs/2604.11748.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Daniel Duckworth, Arvind Neelakantan, Ben Goodrich, Lukasz Kaiser, and Samy Bengio. 2019. Parallel scheduled sampling. *ArXiv*, abs/1906.04331.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *ArXiv*, abs/2210.08933.

Zhen Guo and Fei Tan. 2026. Lost in diffusion: Uncovering hallucination patterns and failure modes in diffusion large language models. *ArXiv*, abs/2604.10556.

Michael Hersche, Samuel Moor-Smith, Thomas Hofmann, and Abbas Rahimi. 2025. Soft-masked diffusion language models. *ArXiv*, abs/2510.17206.

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. 2025. Don't settle too early: Self-reflective remasking for diffusion language models. *arXiv preprint arXiv:2509.23653*.

Jaehyeong Jo and Sung Ju Hwang. 2025. Continuous diffusion model for language modeling. *ArXiv*, abs/2502.11564.

Mingyu Jo, Jaesik Yoon, Justin Deschenaux, Caglar Gulcehre, and Sungjin Ahn. 2025. Loopholing discrete diffusion: Deterministic bypass of the sampling wall. *ArXiv*, abs/2510.19304.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. 2023. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *ArXiv*, abs/2310.02279.

Jaeyeon Kim, Seunggeun Kim, Taekyun Lee, David Z. Pan, Hyeji Kim, Sham M. Kakade, and Sitan Chen. 2025. Fine-tuning masked diffusion for provable self-correction. *ArXiv*, abs/2510.01384.

Barbara Kitchenham, Stuart Charters, and 1 others. 2007. Guidelines for performing systematic literature reviews in software engineering.

Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. Innovations in Software Engineering and Software Development. Chapman and Hall/CRC, Boca Raton, FL.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel F. Shah, Jerry M. Huang, Aditi Raghunathan, Seunghoo Hong, Nicholas Matthew Boffi, and Jinwoo Kim. 2026. Flow map language models: One-step language modeling via continuous denoising.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025. A survey on diffusion language models. *ArXiv*, abs/2508.10875.

Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji rong Wen. 2023. Diffusion models for non-autoregressive text generation: A survey. *ArXiv*, abs/2303.06574.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Ronghua Li, Feiyu Xiong, and Zhiyu Li. 2024. Internal consistency and self-feedback in large language models: A survey. *ArXiv*, abs/2407.14507.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Text diffusion with reinforced conditioning. *ArXiv*, abs/2402.14843.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2023. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Weichao Wang, Xingshan Zeng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Self: Self-evolution with language feedback.

Linrui Ma, Yufei Cui, Kai Han, and Yunhe Wang. 2026. Diffusion in diffusion: Reclaiming global coherence in semi-autoregressive diffusion. *ArXiv*, abs/2601.13599.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Rabeeh Karimi Mahabadi, Jaesung Tae, Hamish Ivison, James Henderson, Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2023. Tess: Text-to-text self-conditioned simplex diffusion. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv*, abs/2305.14251.

Dat Nguyen-Cong, Tung Kieu, and Hoang Thanh-Tung. 2026. Fastdiss: Few-step match many-step diffusion language model on sequence-to-sequence generation-full version. *ArXiv*, abs/2604.05551.

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

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and 1 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Fred Zhangzhi Peng, Zachary Bezemek, Jarrid Rector-Brooks, Shuibai Zhang, Anru R. Zhang, Michael M. Bronstein, Avishek Joey Bose, and Alexander Tong. 2025a. Planner aware path learning in diffusion language models training. *ArXiv*, abs/2509.23405.

Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Alexander Tong, and Pranam Chatterjee. 2025b. Path planning for masked diffusion model sampling. *ArXiv*, abs/2502.03540.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Neural Information Processing Systems*.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. 2025. Diffusion beats autoregressive in data-constrained settings. *ArXiv*, abs/2507.15857.

Yanyu Qian, Yue Tan, Yixin Liu, Wang Yu, and Shirui Pan. 2026. Dynhd: Hallucination detection for diffusion large language models via denoising dynamics deviation learning.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-distribution detection and selective generation for conditional language models.

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. *ArXiv*, abs/2406.07524.

Earl J St Sauver. 2026. Think first, diffuse fast: Improving diffusion language model reasoning via autoregressive plan conditioning. *ArXiv*, abs/2603.13243.

Yair Schiff, Omer Belhasin, Roy Uziel, Guanghan Wang, Marianne Arriola, Gilad Turok, Michael Elad, and Volodymyr Kuleshov. 2026. Learn from your mistakes: Self-correcting masked diffusion models. *ArXiv*, abs/2602.11590.

Yash Shah, Abhijit Chakraborty, Naresh Kumar Devulapally, Vishnu Suresh Lokhande, and Vivek Gupta. 2026a. Oscar: Orchestrated self-verification and cross-path refinement. *ArXiv*, abs/2604.01624.

Yash Shah, Abhijit Chakraborty, and Vivek Gupta. 2026b. Care compute-aware remasking evaluation protocol for masked diffusion language models.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2021. Self-conditioning pre-trained language models. In *International Conference on Machine Learning*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *North American Chapter of the Association for Computational Linguistics*.

Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10.

Zhangkai Wu, Xuhui Fan, Hongyu Wu, and Longbing Cao. 2025. Scot: Unifying consistency models and rectified flows via straight-consistent trajectories. *Advances in Neural Information Processing Systems 38*.

Zhengxian Wu, Kai Shi, Chuanrui Zhang, Zirui Liao, Jun Yang, Ni Yang, Qiuying Peng, Luyuan Zhang, Hangrui Xu, Tianhuang Su, Zhenyu Yang, Haonan Lu, and Haoqian Wang. 2026. When models judge themselves: Unsupervised self-evolution for multimodal reasoning.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. 2025. Conditioning residuals for diffusion models via representation feedback.

Runpeng Yu, Qi Li, and Xinchao Wang. 2025. Discrete diffusion in large language and multimodal models: A survey. *ArXiv*, abs/2506.13759.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Shuibai Zhang, Fred Zhangzhi Peng, Yiheng Zhang, Jin Pan, and Grigorios G. Chrysos. 2025. Corrective diffusion language models. *ArXiv*, abs/2512.15596.

Tunyu Zhang, Xinxi Zhang, Ligong Han, Haizhou Shi, Xiaoxiao He, Zhuowei Li, Hao Wang, Kai Xu, Akash Srivastava, Hao Wang, Vladimir Pavlovic, and Dimitris N. Metaxas. 2026. T3d: Few-step diffusion language models via trajectory self-distillation with direct discriminative optimization. *ArXiv*, abs/2602.12262.

A Library-Specific Search Strings

The canonical Boolean string (§3.2) was adapted per library: arXiv (`all:` field, `cs.CL`/`cs.LG` categories); ACL Anthology (full-text); OpenReview (venue-scoped to ICLR/NeurIPS/ICML/*ACL 2021–2026); IEEE Xplore and ACM DL (metadata + abstract, manual); Semantic Scholar (API, restricted to Computer Science, years 2021–2026). Wildcards ($\star$) were expanded where a library lacked stemming. Forward/backward snowballing used Semantic Scholar and DBLP citation graphs over the studies that passed screening.

Post-corpus context. Works not entering the final systematic corpus because they fall outside Eq. (1) or appeared after the corpus refresh may be discussed as boundary or follow-up references where relevant; they do not alter the screening counts, quality assessment, or corpus-level findings.

B Search Log, Exclusion Breakdown, and Coding Reliability

This appendix releases the audit trail summarised in §3. The complete per-record log (titles, sources, screening decisions), the per-rater quality sheets, and the machine-readable extraction matrix are provided in the anonymised replication repository[2] (`search_log.csv`, `screening_sheet.csv`, `qa_rater1.csv`, `qa_rater2.csv`, `extraction_matrix.csv`); the tables below summarise them.

Per-library hit counts. Table 2 reports imported records per library for the canonical string (§3.2, executed March 2026 in the expanded high-recall configuration). IEEE Xplore and ACM Digital Library were searched manually owing to public API restrictions: IEEE returned no importable records, and ACM's 45 manual hits were checked against the included set but not imported into the automated deduplication pipeline. Cross-library overlap accounts for the reduction from 1,595 imported records to 1,168 unique records after deduplication.

Exclusion breakdown. Title/abstract exclusions are coded by reason in Table 3. The dominant category is database hits that were diffusion- or language-model adjacent but contained no self-feedback or trajectory component, consistent with the deliberately high-recall search. The 729 full-text exclusions were recorded individually during human full-text curation in `screening_sheet.csv` rather than aggregated into categories.

[2] https://github.com/coral-lab-asu/Self_Trajectory_DLM.git

Library	Imported
arXiv (`cs.CL`/`cs.LG`)	635
Semantic Scholar (CS, 2021–2026)	491
OpenReview (ICLR/NeurIPS/ICML/*ACL)	413
ACL Anthology (full-text)	56
IEEE Xplore (manual; none importable)	0
ACM Digital Library (manual; not imported)[†]	(45)
Total imported	**1,595**
Unique after deduplication	**1,168**

Table 2: Per-library imported counts (Figure 1 identification stage). [†]ACM's 45 manual hits were screened against the included set but not imported into the automated deduplication pipeline, so they are excluded from the 1,595 total.

Title/abstract exclusion reason	n
No DLM / trajectory-self-feedback relevance (NR)	290
Outside scope after title/abstract (OOS)	67
Image/audio/video or non-language diffusion (EC1)	58
Total excluded at title/abstract	**415**

Table 3: Coded reasons for title/abstract exclusion, keyed to Figure 1; counts are drawn from the released `screening_sheet.csv`.

Table 3 reports exclusions made from title/abstract information only; full-text exclusions were assessed across the remaining 753 records.

ID	Quality criterion (scored 0/0.5/1)
QA1	Are the aims / research questions clearly stated?
QA2	Is the self-feedback mechanism *g precisely* defined?
QA3	Is there a baseline or ablation isolating the TSF effect?
QA4	Are tasks, datasets, and metrics described reproducibly?
QA5	Is faithfulness / hallucination measured?
QA6	Is correction capacity measured?
QA7	Is the family scope (discrete/cont./flow) stated, and is cross-family generality discussed?
QA8	Are limitations / threats to validity discussed?

Table 4: Quality-assessment checklist applied to all 24 primary studies (§3.5).

Coding reliability. Title/abstract filtering was performed as an automated pre-screen over the 1,168 deduplicated records and is therefore *not*

reported as a human inter-rater reliability statistic. Human reliability is reported for the final 24-study quality-assessment stage, where two reviewers independently scored all 24 studies on the eight criteria: quadratic-weighted agreement averaged κ_w=0.89 (range 0.56–1.00 across criteria), lowest for QA7 (cross-family generality) and QA3 (baseline/ablation isolation); the high agreement on QA5/QA6 reflects their near-constant (predominantly zero) distributions and should be read accordingly. Operator-family assignment, the most debatable extraction field, reached κ=0.71 before the decision rules of Appendix C were formalised and κ=0.89 after re-coding. Agreement was lowest on QA3 — whether a study reports a baseline or ablation that isolates the TSF effect — because judging whether a comparison genuinely isolates the mechanism, rather than confounding it with training regime or step budget, requires interpretive judgement. The QA rubric used one calibration example per level (0/0.5/1) per criterion; the rubric and adjudicated examples are in the released repository. All residual disagreements were reconciled to consensus before the scores reported in Appendix F.

C Operator-Family Decision Rules

Assignment to a TSF operator family (§4) follows the ordered decision procedure below, applied to the self-feedback operator g of Eq. (1). The first matching rule wins; ties are broken by the signal on which the training loss is computed.

Top-down design space. The corpus-derived families can also be recovered by crossing two design axes: *what the feedback acts on* (the denoiser input vs. the step-selection distribution) and *how the fed-back prediction is weighted* (raw, reward-scaled, or error-paired). Raw feedback yields PSC when it enters the denoiser and CPS when it routes step selection; reward scaling yields RRC on the denoiser-input side and an *empty, open RL-routing cell* on the step-selection side; error pairing yields CTF on the denoiser-input side and a hybrid routing/correction cell exemplified by RemeDi-like operators. Empty cells, new weighting modes, and genuinely co-equal multi-channel operators therefore extend the design space rather than being forced into an existing family. The ordered assignment rules below and Appendix C operationalise this mapping for the reviewed corpus.

D1 Does g feed back the *full* previous prediction $\tilde{x}_{t+1}$ (one-hot, simplex, or embedding) as denoiser input, with no reward or explicit error channel? ⇒ **PSC**. If the fed-back signal is a confidence-weighted superposition designed to avoid hard commitment ⇒ **PSC sub-type (soft-masking)**.

D2 Does g convert the prediction into a distribution over token *positions* used to route or plan unmasking, rather than as denoiser input? ⇒ **CPS**.

D3 Does g scale the fed-back signal by a reward/advantage estimate? ⇒ **RRC**.

D4 Does g pass an explicit error signal e_{t+1} flagging likely-wrong positions alongside the prediction? ⇒ **CTF**.

D0 If none applies, the study uses **no** TSF (default denoising).

Table 5 records the cases that triggered rater disagreement and their adjudication; these are the cases that motivated D1–D4 and drove the post-formalisation agreement from κ=0.71 to 0.89 (Appendix B).

Study	Candidates	Resolution (rule)
Soft-Masked	PSC vs sub-type	Sub-type: superposition avoids hard commitment but keeps PSC operator structure (D1)
P2	CPS vs PSC	CPS: confidence routes *positions*, not used as denoiser input (D2)
TREC	RRC vs PSC	RRC: feedback is reward-scaled, not the raw prediction (D3)
SCMDM	PSC vs CTF	PSC: two-pass self-conditioning, no explicit error channel (D1)
PAPL	CPS vs none	CPS: planner weighting is confidence-derived routing (D2)

Table 5: Adjudicated borderline operator-family assignments.

Hybrid and multi-channel operators. Some operators combine channels. RemeDi (Huang et al., 2025) jointly predicts a token distribution and a per-token confidence score and remasks low-confidence tokens: the confidence head routes which positions are revisited (CPS-like) while remasking targets likely-wrong tokens (CTF-like). We assign such operators by their *primary* channel which is the signal on which the training loss is computed, while recording auxiliary channels

as annotations; an operator that genuinely fuses two channels with co-equal losses extends the taxonomy as an ordered (primary, auxiliary) pair to which D1–D4 apply. We cite RemeDi here only to illustrate a hybrid boundary: it is not one of the 24 primary-study rows in Table 8, so we do not assign it a QA score or use it in corpus-level denominators.

D SEP-DLM Reference Feasibility Run

The rebuttal reference run uses LLaDA-3B (Nie et al., 2025) on $N{=}400$ PopQA examples (Mallen et al., 2023). The decoder's remasking/unmasking schedule is shared by the strength-0 and strength-1 endpoints; the manipulated quantity is feedback strength. This controls the decoder schedule within the pilot, but does not make the result a general causal claim across models or datasets.

Metric	TSF on	TSF off	Δ
Generative PPL ↓	164.9	610.3	−445.4
Factual accuracy ↑	0.220	0.098	+0.123
Token-F1 ↑	0.127	0.041	+0.086
Hallucination rate ↓	0.780	0.903	−0.123

Table 6: Single-backbone SEP-DLM feasibility run (LLaDA-3B/PopQA, $N{=}400$). Values are point estimates and are not presented as a validated benchmark or significance result.

Feedback strength	0.00	0.25	0.50	0.75	1.00
Accuracy	0.090	0.175	0.180	0.200	0.210
Hallucination rate	0.910	0.825	0.820	0.800	0.790

Table 7: Feedback-strength sweep with the decoder schedule held fixed. The monotone pattern is suggestive but is not used as a substitute for statistical uncertainty.

Diagnostic instantiation. The run instantiates Commitment Index with a hard-flip intervention (force a wrong token and measure subsequent revision), yielding hard-flip AUC ≈ 0.12. A σ sweep found CI nearly flat over the tested perturbation magnitudes, which is why we remove the earlier universal 0.5-nat recommendation. Correction Capacity uses lexical token swaps and type-matched semantic substitutions from the knowledge-base distractor set, yielding $\mathrm{CCap} \approx 0$ for lexical and $\mathrm{CCap} \approx 0.13$ for semantic corruption. Code and per-axis logs are released alongside the review artifacts.

Uncertainty and scope. At the end of the rebuttal period, bootstrap confidence intervals for these pilot outputs had not been finalised. We therefore preserve the run as a check of end-to-end feasibility and report no formal significance claim. The one-backbone, one-dataset, $N{=}400$ design is insufficient to validate CI, CCap, or SEP-DLM outside this setting; replication with confidence intervals and controlled cross-family runs remains part of the research agenda.

E Diagnostic Protocols

This appendix specifies the implementation choices that make the Commitment Index (CI, §7) and Correction Capacity (CCap) reproducible across labs; a reference implementation with these defaults and fixed seeds is released.

CI **operationalisation.** Three choices must be reported. (i) *Intervention and perturbation.* We do not prescribe a universal KL target. In a discrete MDLM, the reference feasibility run uses a hard-flip intervention—force an incorrect token at the chosen step and measure whether the later trajectory revises it—together with a reported σ sweep in the native pre-softmax signal space. CI was nearly flat across the tested σ settings, so the rebuttal result does not support the earlier arbitrary 0.5-nat recommendation. Continuous/flow studies should analogously calibrate and report a family-appropriate perturbation sweep rather than inherit a fixed constant. (ii) D_{edit} is the token-level Levenshtein distance after greedy decode, normalised by $\max(|y_{\mathrm{ref}}|, |y|)$ so that $D_{\mathrm{edit}} \in [0, 1]$; sequences are length-aligned by the standard edit-distance alignment, not by position. (iii) *Toggling* ablates only the self-feedback channel s_{t_0} (Eq. (1)) while leaving x_{t_0}, context c, and all other steps untouched, so that CI reflects the intended intervention rather than unrelated decoder changes. Future validation runs should report $\mathrm{CI}(t)$ with bootstrap confidence intervals over paired examples/seeds.

Cross-family use. The reference run below instantiates CI on one discrete backbone, but it does not test $\mathrm{H}_{\mathrm{wall}}$. A future matched-compute comparison of discrete and continuous/flow models would test the hypothesis directly: earlier commitment in the discrete model would be consistent with $\mathrm{H}_{\mathrm{wall}}$, whereas the opposite ordering would falsify that prediction.

CCap **corruption protocol.** Comparability requires fixing the corruption process, the ground-truth alignment, and the uncertainty estimate. Report CCap under two corruption families separately, as they probe different repair abilities: *lexical* corruption (random token substitutions/deletions at rate ϵ; the pilot uses token swaps, diagnosing surface-level repair) and *semantic* corruption (replacing a span with a plausible but factually wrong, type-matched alternative from the underlying knowledge-base distractor set, diagnosing knowledge-grounded repair). Positions are aligned to $x^\star$ by the same edit-distance alignment used for CI; a position counts as repaired only if it matches $x^\star$ after K steps and as regressed only if a position that matched $x^\star$ no longer does. Report CCap as a curve over $\epsilon \in \{0.05, 0.1, 0.2\}$ and budgets $K \in \{4, 8, 16\}$ with bootstrap confidence intervals over seeds, so repair behaviour reads as a function of corruption severity and compute rather than a single point estimate.

Compute accounting. To prevent iso-compute disputes, report effective decoding cost as (forward passes per step) $\times$ (steps T) $\times$ (per-pass FLOPs), and compare methods at matched totals: two-pass PSC/CTF variants are compared at $T/2$ steps (or the baseline re-run at $2T$), planner-based CPS counts its per-step routing cost, and flow/one-step models report passes for their actual schedule (e.g. $T{=}1$ for distilled maps). Because FLOPs miss some bottlenecks (e.g. memory bandwidth), also report wall-clock and peak memory, and count any cached or partially-reused pass by its reused fraction rather than as a full second pass. A reporting checklist is in Appendix E. Recent post-corpus evaluation work reinforces this need: CaRE (Shah et al., 2026b) standardises actual numbers of function evaluations (NFE), multi-metric reporting, and sampling stochasticity for masked-DLM remasking comparisons, complementing SEP-DLM's broader quality, correction, commitment, and faithfulness axes.

Rationale for the axes. Each axis targets a failure mode the synthesis surfaced: *quality* pairs GenPPL with MAUVE (the premature-termination and degeneration modes of C5, Chang et al., 2025; Guo and Tan, 2026); *correction* (CCap) operationalises C4; *commitment* (CI) targets early lock-in; and *faithfulness* (hallucination rate, FactScore) is the axis QA5 shows is never measured while varying TSF (RQ5). The axes are orthogonal, so a gain on one cannot mask a regression on another, with recommended metrics and benchmarks in Table 11 (Appendix E).

Reporting checklist. A study claiming a TSF effect should report, under SEP-DLM: (1) the operator family (PSC/CPS/RRC/CTF/none) and the primary/auxiliary channels of g; (2) all four axes — quality (GenPPL, MAUVE), correction (CCap), commitment (CI), and faithfulness (hallucination rate, FactScore); (3) the three compute-accounting numbers (passes/step $\times T\times$ per-pass FLOPs), plus wall-clock and peak memory and how cached/partial passes are counted; and (4) seeds, tokenizer, and, for multilingual runs, per-script distractors. The released scaffold instantiates these defaults.

Multilingual and cross-domain. CI and CCap are language-sensitive. For non-English settings we recommend a fixed per-script subword tokenizer, an embedding-based semantic delta reported alongside CI's Levenshtein distance (so paraphrastic edits are not over-penalised), and language-specific distractor sets for CCap's semantic corruption rather than translated English ones. SEP-DLM should thus specify a multilingual slice; the same re-calibration applies to domain shift (code, specialist text), lest the protocol entrench the English-centric monoculture flagged in Limitations.

F Per-Study Extraction Matrix and QA Scores

Table 8 expands the cluster matrix to one row per primary study, with the seven extraction fields condensed and the QA total (max 8). QA5 (faithfulness measured) and QA6 (correction measured) are reported explicitly because their near-zero column sums constitute the review's central finding. Table 10 separates text/knowledge, code/math, and non-text evidence for RQ5.

Faithfulness-boundary sensitivity. The numerical intersection depends on what is called "faithfulness." Under the *Strict* definition used for RQ5 (factual accuracy, hallucination rate, FactScore, or equivalent knowledge-grounded factual metrics), the intersection is **0/24**. Under a *Broad* definition that additionally admits functional correctness such as `pass@k` and protein foldability, it becomes **1/24** (P2). Under a *Broadest* definition that also admits PPL-family quality proxies

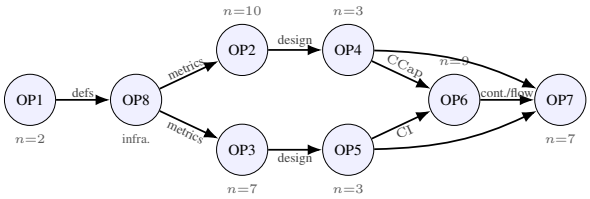


Figure 2: Open-problem dependency map (§6). Edges denote "should be resolved before" and are labelled with the bridge each precedence supplies. Node annotations give the number of primary studies whose extracted open problems point to each node; OP8 is corpus-wide (marked *infra.*). OP2 (n=10) and OP6 (n=9) are most heavily implicated, both gated by OP1 and OP8. Each edge implies a concrete test under SEP-DLM: e.g. OP2→OP4, run CCap under PSC/RRC/CTF at $K \in \{4, 8, 16\}$ and $\epsilon \in \{0.05, 0.1, 0.2\}$ at iso-compute; OP4/OP5→OP6, add a faithfulness axis to those runs; OP6→OP7, repeat on a continuous/flow backbone to test $\mathrm{H}_{\mathrm{wall}}$.

(PPL/GenPPL/perplexity), it becomes **8/24**. Thus the zero-count headline is specifically a factual-faithfulness result; broadening the outcome definition produces non-zero overlaps and should not be conflated with factual faithfulness. Appendix Table 10 separates text/knowledge, code/math, and non-text evidence.

Coverage snapshot. The denominator is auditable by sub-field and venue: 16 of 24 studies vary a TSF mechanism and 8 vary none (families/clusters in Table 8), 22 of 24 are arXiv-indexed, and 15 from 2025–2026 are the preprint reliance and recency that motivate bounding the claim to our search window (Appendix H).

Cross-cluster tensions and complementarities. Read together, the clusters expose three disconnects that are obscured by paper-by-paper summaries. First, C2/C3 optimise quality, efficiency, or path selection, whereas C5 measures factual failure; evidence that a trajectory is *better scored* is therefore not evidence that it is more faithful. Second, C4 establishes that targeted correction is trainable, but almost entirely on discrete code/math or synthetic corruption, while C5 observes naturalistic hallucination without a TSF intervention. Third, C6/C7 broaden the architectural space but rarely measure correction or faithfulness. These are complementary rather than contradictory literatures: together they motivate controlled experiments that connect mechanism, trajectory behaviour, and factual outcomes under a common protocol.

G Related Survey Comparison

Table 12 compares this survey against five adjacent prior surveys on the dimensions of scope, method, DLM coverage, TSF mechanism coverage, systematic search, and faithfulness gap analysis; each row states the relevant scope boundary. Referenced from §8.

H Threats to Validity

Following Kitchenham et al. (2015), by category. **Construct:** the four-family mapping is interpretive, mitigated by explicit decision rules (Appendix C) and double-coding of operator family. **Internal:** a six-library string plus one snowballing iteration may miss recent preprints; we used recall-oriented terms, a 2021–2026 database-search window, a pre-submission forward-snowballing refresh (adding SCMDM), and a five-paper known-item check that recovered 5/5 seeds. The seed check is deliberately small and should not be read as a comprehensive recall estimate. **External:** findings concern small-to-medium DLMs, and several continuous/flow checkpoints lack public implementations. **Conclusion:** $\mathrm{H}_{\mathrm{wall}}$ is untested and possibly confounded by architecture, data, or scale; the diagnostics and SEP-DLM have one feasibility instantiation but are not broadly validated, and the CI perturbation/intervention must be calibrated and reported per family. The pilot itself is limited to LLaDA-3B, PopQA, N=400, and point estimates without final bootstrap intervals. **Reporting bias:** original-paper numbers do not control for corpus, tokeniser, or step budget, so the gap matrix is *directional*, hence SEP-DLM's controlled conditions.

Study	Cluster	TSF form	Key metrics	QA5	QA6	QA tot.	OP(s)
D3PM (Austin et al., 2021)	C1	None	PPL	0	0	5.0	OP2
DiffuSeq (Gong et al., 2022)	C1	None	BLEU, diversity	0	0	5.0	OP2,OP3
SEDD (Lou et al., 2023)	C1	None	PPL	0	0	5.5	OP2
MDLM (Sahoo et al., 2024)	C1	None (declined)	PPL	0	0	6.0	OP2,OP3
LLaDA (Nie et al., 2025)	C1	Optional remask	NLU acc.	0	0	5.5	OP2,OP7
Analog Bits (Chen et al., 2022)	C2	PSC	FID/quality	0	0	5.5	OP1,OP2
TESS (Mahabadi et al., 2023)	C2	PSC (simplex)	BLEU, ROUGE	0	0	5.5	OP1,OP6
TREC (Liu et al., 2024)	C2	RRC	BLEU, MAUVE	0	0	6.0	OP3
FastDiSS (Nguyen-Cong et al., 2026)	C2	PSC (few-step)	GenPPL, speed	0	0	5.5	OP2,OP3
Soft-Masked (Hersche et al., 2025)	C2	PSC sub-type	PPL (iso-compute)	0	0	5.5	OP2,OP7
SCMDM (Cardei et al., 2026)	C2	PSC (post-train)	GenPPL (OWT)	0	0	6.0	OP2,OP3
P2 (Peng et al., 2025b)	C3	CPS	`pass@k`, fold.	0	0	6.0	OP3,OP6
PAPL (Peng et al., 2025a)	C3	CPS (P-ELBO)	MAUVE, GenPPL	0	0	6.0	OP3,OP6
CDLM (Zhang et al., 2025)	C4	CTF	Code revision (CRB)	0	1	6.5	OP4,OP6
ProSeCo (Schiff et al., 2026)	C4	CTF	GSM8K, code	0	1	6.5	OP4,OP6
PRISM (Kim et al., 2025)	C4	CTF	Provable corr.	0	1	6.0	OP4,OP6
TraceDet (Chang et al., 2025)	C5	None varied	AUROC	1	0	6.5	OP5,OP6
DynHD (Qian et al., 2026)	C5	None varied	AUROC, entropy	1	0	6.5	OP5,OP6
LostInDiff. (Guo and Tan, 2026)	C5	None varied	Halluc. rate	1	0	6.5	OP5,OP6
LangFlow (Chen et al., 2026)	C6	PSC	PPL, MAUVE	0	0	6.0	OP7
RDLM (Jo and Hwang, 2025)	C6	PSC	PPL	0	0	5.5	OP7
FLM/FMLM (Lee et al., 2026)	C6	Flow map	PPL, speed	0	0	5.0	OP7
DLM-One (Chen et al., 2025)	C6	Flow map	PPL, 1-step speed	0	0	5.0	OP7
LDDM (Jo et al., 2025)	C7	PSC / latent	PPL	0	0	5.5	OP2,OP7
			Column sums (QA5 / QA6):	**3**	**3**	*max 6.5*	

Table 8: Per-study extraction matrix and quality scores (Appendix F); the full QA1–QA8 breakdown is released as `extraction_matrix.csv`. Across all 24 studies, faithfulness (QA5) is measured by only the three hallucination studies—*none of which varies the TSF mechanism*—and correction (QA6) only by the three CTF studies, *none of which measures faithfulness*. Because no study scores on both QA5 and QA6, no included study exceeds $7/8$: the missing point is the review's central gap (OP6) made numerical. The empty intersection of QA5 and "TSF form varied" is the review's central finding: every study scoring QA5=1 (TraceDet, DynHD, LostInDiff.) carries TSF form "None varied," so faithfulness reporting and TSF variation never co-occur in a single row.

Cluster	Studies	TSF form	Measured	Hidden assumption	Blind spot	OP
C1 Foundational	D3PM, DiffuSeq, SEDD, MDLM, LLaDA	None / optional remask	PPL, quality	Independent-step optimality suffices	No correction or faithfulness	OP2, OP3
C2 Self-cond. & eff.	Analog Bits, TESS, TREC, FastDiSS, Soft-Masked, SCMDM	PSC / RRC	PPL, BLEU, MAUVE, GenPPL	Quality gains imply reliable feedback	No hallucination or correction	OP1–OP3, OP6
C3 Planning	P2, PAPL	CPS	`pass@k`, fold., MAUVE	Better paths imply better outputs	No faithfulness analysis	OP3, OP6
C4 Correction	CDLM, ProSeCo, PRISM	CTF	Code/math correction	Synthetic correction transfers broadly	No naturalistic halluc.; discrete only	OP4, OP6
C5 Hallucination	TraceDet, DynHD, LostInDiff.	None varied	AUROC, halluc. rate	Default trajectories reveal all failures	No TSF intervention	OP5, OP6
C6 Cont./flow	LangFlow, RDLM, FLM, DLM-One	PSC / flow map	PPL, quality, speed	Quality metrics capture reliability	No CCap/CI/halluc.; no discrete comp.	OP7
C7 Wall mitigation	LDDM, Soft-Masked	PSC subtype / latent	PPL, quality	Removing commitment $\Rightarrow$ better gen.	No faithfulness/correction test	OP2, OP7

Table 9: Cluster-level gap matrix (RQ2–RQ5), expanding the cluster narrative of §5. The per-study expansion (24 rows, with QA scores) is Table 8.

Panel	TSF family/form	Evidence in the reviewed corpus	Interpretation for the faithfulness gap
A. Text / knowledge			
	PSC	TESS, SCMDM, LangFlow: BLEU/ROUGE/PPL/MAUVE/GenPPL	Quality evidence; no knowledge-grounded factual-faithfulness evaluation while varying PSC.
	CPS	PAPL: MAUVE/GenPPL	Path/quality evidence; no factual-faithfulness evaluation.
	RRC	TREC: BLEU/MAUVE	Reward-scaled feedback evidence; no factual-faithfulness evaluation.
	None varied	TraceDet, DynHD, Lost in Diffusion: AUROC/hallucination	Faithfulness is measured, but the TSF mechanism is not varied.
B. Code / math			
	CPS	P2: code `pass@k`	Functional correctness is adjacent to, but distinct from, the strict factual-faithfulness definition.
	CTF	CDLM, ProSeCo, PRISM: code revision/GSM8K/provable correction	Correction is measured under code/math or synthetic settings, not naturalistic factual QA.
C. Non-text			
	CPS	P2: protein foldability	Non-text functional evidence is reported separately and is not pooled with factual-faithfulness outcomes.

Table 10: Domain-separated evidence for RQ5. Separating text/knowledge, code/math, and non-text outcomes prevents heterogeneous metrics from being treated as directly comparable. P2 appears in Panels B and C because it reports both code `pass@k` and protein foldability.

Axis	Metric	Benchmark
Generation quality	GenPPL, MAUVE	OpenWebText, WMT
Correction capacity	$\mathrm{CCap}(\epsilon, K)$	CRB, $\mathrm{GSM8K}_{\mathrm{corrupt}}$, $\mathrm{TriviaQA}_{\mathrm{corrupt}}$
Traj. commitment	$\mathrm{CI}(t)$ curve	Paired bifurcation trajectories
Faithfulness	Halluc. rate, FactScore	TriviaQA(Joshi et al., 2017), PopQA(Mallen et al., 2023), FEVER(Thorne et al., 2018)

Table 11: SEP-DLM: four axes recommended for any DLM self-feedback claim (RQ6), expanding §7 (Review-Derived Evaluation Recommendations).

Work	Scope	Method	DLM coverage	TSF mechanism coverage	Systematic search	Faithfulness gap analysis
Li et al. (2023)	Diffusion models for non-AR text generation	Narrative survey	Broad text diffusion	Not TSF-specific	No TSF-specific denominator	No
Li et al. (2025)	DLM architecture, training, inference	Narrative survey	Discrete, continuous, flow	Mentioned per model	No auditable denominator	No
Yu et al. (2025)	Discrete DLMs	Narrative survey	Discrete only	Partial	No auditable denominator	No
Pan et al. (2024)	Automated LLM self-correction	Narrative survey	AR/post-hoc boundary	Self-correction, not DLM TSF	No DLM-TSF denominator	No
Liang et al. (2024)	AR self-feedback	Survey	Boundary only	AR feedback, not DLM TSF	Not DLM-focused	No
This survey	TSF in DLMs	Systematic review	Discrete, continuous, flow	Four-family taxonomy	Yes	Yes

Table 12: Comparison with related surveys and explicit scope decisions. Readers seeking broad diffusion/non-autoregressive text coverage should consult Li et al. (2023) and Li et al. (2025); readers seeking discrete-DLM depth should consult Yu et al. (2025); readers studying automated AR self-correction or AR self-feedback should consult Pan et al. (2024) and Liang et al. (2024). This survey is vertical rather than horizontal: it isolates TSF mechanisms in DLMs with an auditable denominator and factual-faithfulness gap analysis.