WORDS, NOT WALKS: AN ALGEBRAIC THEORY OF TOOL-USE TRAJECTORY REDUCTION FOR AGENT DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A language agent that solves a task by calling tools produces a *trajectory*: a sequence of actions that transforms an environment state. We argue that this process is most naturally described not as a walk on a graph but as a *word* in a transformation monoid: tools are generators, dependencies are non-commutativity, and an efficient solution is a *geodesic*—a shortest generator product. This view exposes two failure modes that plague behavioral cloning of agents: state-relative identity sub-words (no-ops and backtracking loops that fix the current state) and non-commutative ordering errors. We introduce *Algebraic Trajectory Reduction* (ATR), a sound, model-free procedure that rewrites a teacher trajectory to its empirical geodesic before it is used for supervised fine-tuning. ATR unifies loop elimination and observed shared-state shortcuts as shortest-path extraction over the observed transition graph, with provable soundness under deterministic tools and exact state identifiers. We further give a sufficient boundary: on strictly monotone response integration, ATR coincides with deduplication by construction, whereas state revisits in sequential tool-use can make dedup unsound. Both regimes appear in real data. On $44{,}730$ GPT-4/DFSDT TOOLBENCH trajectories, 27.9% of API calls fail (no-information updates) and ATR removes 39.4% of calls, exactly matching the strongest baseline as predicted. On 336 sequential AGENTINSTRUCT/ALFWorld trajectories, ATR removes twice as much as dedup (24.4% vs. 12.0%); under the observation proxy, offline replay fails on 1.5% of ATR paths versus 25.3% for dedup, and the former 1.5% is already present in unreduced data. A conservative certified variant removes only 1.3%. A 3B Qwen student trained on the proxy-reduced supervision achieves closed-loop success comparable to RAW/B2 within seed variation (9.5% vs. $9.2\%/10.0\%$) using 24.4% fewer action targets. In a controlled transformation-monoid environment where teacher verbosity is a manipulated variable, cloning identical policies on ATR-reduced data raises closed-loop task success from 71.4% to 97.2% (matching an oracle-geodesic skyline of 97.8%), cuts non-commutativity errors from 40.9% to 0.7%, and uses 46% fewer supervision transitions; ATR is invariant to how verbose the teacher is, whereas naive cloning degrades monotonically. Our results give a precise algebraic account of *what makes an agent trajectory worth imitating*.

1 INTRODUCTION

Turning a strong but expensive model into a cheap, capable agent is usually done by *trajectory distillation*: collect successful tool-use trajectories from a teacher (e.g. GPT-4) and supervised-fine-tune (SFT) a smaller student on them (Qin et al., 2024; Zeng et al., 2024; Chen et al., 2023). A recurring, and under-theorized, difficulty is that *not all successful trajectories are equally worth imitating*. Teacher trajectories frequently contain redundant steps—probes that change nothing, walks that return to a previous situation, tentative actions that are immediately undone—and a student that clones them inherits these pathologies, producing longer, looping, and precondition-violating behavior (Shinn et al., 2023; Liu et al., 2024; Chen et al., 2024).

We take the position that the right object of study here is *algebraic*. A tool is a transformation of the environment state, defaulting to a no-op when its execution precondition fails; the set of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tools generates a *transformation monoid* acting on the state space (Howie, 1995); a trajectory is a *word* (a product of generators); and a task is solved by any word carrying the initial state to the goal. Under this lens (i) tool *dependencies* manifest as *non-commutativity* ($t_b \circ t_a \neq t_a \circ t_b$): `take` does not commute with `go to`; (ii) *no-ops and dead loops* are state-relative identity sub-words ($u(s) = s$); and (iii) an *efficient* solution is a *geodesic*, the shortest generator product reaching the goal—a geodesic-reachability problem on the induced action graph, related to geodesic words in geometric group theory (Epstein et al., 1992). To our knowledge, this standard algebraic vocabulary has not been used to formalize and repair agent tool-use, even though many recent methods implement algebraic operations under other names (tree search for geodesics (Zhou et al., 2024a; Besta et al., 2024), heuristic trajectory pruning (Qin et al., 2024; Chen et al., 2024), and step-penalized preference optimization (Putta et al., 2024)).

Building on this formalization we introduce *Algebraic Trajectory Reduction* (ATR): a procedure that rewrites a teacher trajectory into a behaviorally-equivalent, shorter word before SFT. ATR can be understood locally as loop-eliminating rewriting, but is operationally defined by shortest-path extraction over the *observed* state-transition graph. We prove that loop rewriting is sound and terminating and that graph extraction is shortest over observed paths (Section 4). Crucially, ATR is *model-free*: it needs only the state identifiers already present in the trajectory, not a simulator or reward model.

We validate the theory on both real and controlled data. **(1) Real data** (Section 5): we analyze two regimes. On $44{,}730$ GPT-4/DFSDT TOOLBENCH trajectories (Qin et al., 2024)—the search-collected data TOOLLLAMA is trained on—27.9% of API calls fail (no-information updates) and 14.3% of successive calls have a *query-construction* dependency (a later call's arguments come from an earlier response). At the realized-response integration layer used for reduction, successful look-ups accumulate into an order-independent information set; this quotient is distinct from call construction, and ATR here collapses to the best baseline, removing 39.4% of calls. On 336 sequential AGENTINSTRUCT/ALFWorld trajectories (Zeng et al., 2024), every successful `take` is enabled by a preceding `go to` and 93.7% of `puts` occur while holding—a non-Abelian action monoid—and here ATR removes twice as much as deduplication (24.4% vs. 12.0%), with 1.5% versus 25.3% offline replay failures under the observation proxy. **(2) Controlled study** (Section 6): in a transformation-monoid "Delivery Workshop" where we *manipulate* teacher verbosity and compute true geodesics, we clone identical policies on RAW (verbose teacher), ATR-reduced, and ORACLE-geodesic supervision. ATR lifts closed-loop success from 71.4% to 97.2%— statistically indistinguishable from the ORACLE skyline (97.8%)—collapses precondition-violating (non-commutativity) errors from 40.9% to 0.7%, and does so with 46% fewer supervision transitions. A redundancy sweep shows ATR is *invariant* to how verbose the teacher is, whereas naive cloning collapses.

Contributions.

- A precise algebraic formalization of tool use as words in a transformation monoid, with a taxonomy of trajectory redundancy (identity sub-words, non-commutative detours) and a geodesic notion of optimality (Section 3).
- ATR, a sound, terminating, model-free reduction with a local rewriting interpretation and an observed-graph shortest-path algorithm; plus a coverage condition for recovering the true geodesic and a sufficient boundary showing ATR equals dedup under strictly monotone information integration but can preserve reachability where dedup fails on sequential tool-use (Section 4).
- Real-data evidence spanning both regimes: the predicted structure (failed-call identities, query-construction dependencies, and—in the sequential regime—true non-commutative detours) is pervasive in TOOLBENCH and ALFWORLD, and on sequential data ATR is measurably more effective and sounder than the strongest baseline (Section 5). Closed-loop distillation of QWEN2.5-1.5B/3B students shows that ATR's 24.4% shorter supervision causes no detectable 3B degradation within the present three-seed evaluation; at 1.5B it also yields the lowest inadmissible-action and no-op rates (Table 2).
- A controlled causal study isolating supervision quality, showing large and low-variance gains from ATR that match an oracle skyline while being model-free, plus scaling and robustness analyses (Section 6).

2 Related Work

Agent trajectory distillation. SFT on teacher trajectories underlies ToolLLM (Qin et al., 2024), AgentTuning (Zeng et al., 2024), FireAct (Chen et al., 2023), Lumos (Yin et al., 2024), and Toolformer (Schick et al., 2023); these systems build on the thought–action–observation interaction pattern exemplified by ReAct (Yao et al., 2023b). Agent-FLAN (Chen et al., 2024) observes that raw teacher data injects hallucinated and redundant steps and reconstructs the data with hand-designed heuristics. We give this clean-up a mathematical foundation: redundancy is precisely a non-geodesic word, and its removal is geodesic extraction. Our monotonicity condition (Proposition 4) explains why deduplication is optimal for response-set integration and why it can be unsound once sequential trajectories revisit states.

Search and planning for agents. ToolLLM's DFSDT (Qin et al., 2024), LATS (Zhou et al., 2024a), Tree/Graph-of-Thoughts (Yao et al., 2023a; Besta et al., 2024), and Agent Q (Putta et al., 2024) search a state tree/graph for high-reward, compact trajectories—operationally, geodesics on a Cayley-like graph. These methods require a simulator or value model at *search* time; ATR instead *post-processes* already-collected trajectories using only observed states, and is complementary to them. Consequently, our direct baselines share the same fixed trajectories and information budget; comparing end-to-end against online reflection or preference pipelines would additionally vary environment calls, search, and optimization objectives.

Formal models of agents. CoALA (Sumers et al., 2024) casts agents as transition systems; LLM+P (Liu et al., 2023) compiles to PDDL, whose preconditions/effects (Fikes & Nilsson, 1971) are the logical face of non-commutativity. We adopt the dual, *algebraic* face (generators and relations), which makes redundancy and optimality first-class and directly actionable on trajectories. Failure taxonomies in AgentBench (Liu et al., 2024) and Reflexion (Shinn et al., 2023) (action loops, dead loops) are state-relative identity sub-words in our formalism. Our reduction draws on geodesic-word methods in geometric group theory (Epstein et al., 1992) and term rewriting (Baader & Nipkow, 1998); distillation and imitation connections are to Hinton et al. (2015) and DAgger (Ross et al., 2011).

3 The Algebra of Tool Use

States, generators, words. Let S be a set of environment states and $\mathcal{G} = \{t_1, \dots, t_n\}$ a set of tools, each a total map $t_i : S \to S$; outside its execution precondition, $t_i(s) = s$. The tools generate a *transformation monoid* $\mathcal{M} = \langle \mathcal{G} \rangle$ under composition, with identity e (the empty action). A *trajectory* (or *word*) is a finite chronological sequence $W = (a_1, \dots, a_k)$, $a_j \in \mathcal{G}$; its semantics is $W(s) = a_k(\cdots a_1(s))$ and it induces a *walk* $s_0 \xrightarrow{a_1} s_1 \cdots \xrightarrow{a_k} s_k$. Most tools are irreversible (deleting a file, sending a request), so $\mathcal{M}$ is a monoid, not a group; we work with monoids throughout (Howie, 1995).

Definition 1 (Task and solution)**.** *A task is a pair* $(s_{\text{init}}, s_{\text{goal}})$. *A word* W solves *it if* $W(s_{\text{init}}) = s_{\text{goal}}$. *For predicate goals, all guarantees target a chosen observed goal-satisfying endpoint; selecting an earlier satisfying state is a separate stopping decision.*

Non-commutativity is dependency. Two generators *commute at* s if $(t_b \circ t_a)(s) = (t_a \circ t_b)(s)$. An execution *dependency* can be witnessed by a failure of commutation: if `take x` requires being at `x`'s location, then with $t_a =$`go to y`, $t_b =$`take x`, we have $(t_b \circ t_a)(s) \neq (t_a \circ t_b)(s)$—one order succeeds, the other is a no-op. Thus a non-Abelian action monoid captures ordering dependencies. We verify this structure in real data (Section 5).

Definition 2 (Behavioral equivalence and geodesic)**.** *Words* W, W' *are* equivalent on s, $W \sim_s W'$, *if* $W(s) = W'(s)$. *The* geodesic length *of a task is* $d(s_{\text{init}}, s_{\text{goal}}) = \min\{|W| : W(s_{\text{init}}) = s_{\text{goal}}\}$, *the shortest generator product solving it—a geodesic-reachability problem on the action graph (Epstein et al., 1992). A solving word is* redundant *if* $|W| > d(s_{\text{init}}, s_{\text{goal}})$.

A taxonomy of redundancy. Given a solving walk $s_0, \dots, s_k$, redundancy takes two forms. *(i) State-relative identity sub-words.* A contiguous sub-word $U = (a_{j+1}, \dots, a_m)$ with $s_j = s_m$ satisfies $U \sim_{s_j} e$: it lies in the point stabilizer of s_j, though it need not equal e globally, and

Algorithm 1 Algebraic Trajectory Reduction (empirical geodesic)

Require: walk states $s_0, \ldots, s_k$, actions $a_1, \ldots, a_k$
1: build G_W: for $i = 0, \ldots, k-1$, add $s_i \to s_{i+1}$ labeled a_{i+1} (skip self-loops)
2: BFS from s_0; recover a shortest path $s_0 \to s_k$
3: **return** action labels along that path

can be excised. No-ops ($s_j = s_{j+1}$, e.g. a failed precondition or a pure probe) and inverse pairs (`open;close`) are the length-1 and length-2 cases; backtracks and dead loops (Shinn et al., 2023; Liu et al., 2024) are the general case. *(ii) Non-commutative detours.* Even loop-free, an order may be sub-optimal (visiting stations in a poor sequence), so the walk is longer than the geodesic though it revisits no state. ATR eliminates (i) exactly and (ii) to the extent it is exposed by the observed graph.

4 ALGEBRAIC TRAJECTORY REDUCTION

ATR takes a solving trajectory (a walk with state identifiers) and returns a shorter solving word. We give a local rewriting interpretation and the stronger operational shortest-path algorithm.

View 1: term rewriting. Treat the walk as a word and repeatedly apply the length-decreasing rules

$$\begin{aligned} \text{(R1) identity:} \quad & s_j = s_{j+1} \;\Rightarrow\; \text{delete } a_{j+1}; \\ \text{(R2) cancellation:} \quad & s_j = s_{j+2} \;\Rightarrow\; \text{delete } a_{j+1}, a_{j+2}; \\ \text{(R}\ell\text{) loop:} \quad & s_j = s_{j+\ell} \;\Rightarrow\; \text{delete } a_{j+1}, \ldots, a_{j+\ell}. \end{aligned}$$

Each rule removes a sub-word behaviorally equivalent to e at the current state.

View 2: empirical geodesic. Build the directed graph G_W whose vertices are the states visited by W and whose edges are the observed transitions $s_i \xrightarrow{a_{i+1}} s_{i+1}$. Return the labels along a shortest path from s_0 to s_k in G_W (Algorithm 1). Because every edge is a real observed transition, the output is a valid solving word; being a shortest path, it is the geodesic *attainable from observed transitions*. It can exploit a shared-state shortcut only when the requisite intermediate states and edges were actually observed, within one trajectory or a pooled graph.

Proposition 1 (Soundness)**.** *For deterministic tools and exact state identifiers, the word W' returned by Algorithm 1 satisfies $W'(s_0) = s_k$; if W solved the task, so does W'.*

Proof. Every edge of G_W is an observed transition $s \xrightarrow{t} s'$ with $t(s) = s'$. A path $s_0 \to s_k$ is a composition of such transitions, hence $W'(s_0) = s_k$. The rewriting view is a special case: each rule deletes a sub-word U with $U(s_j) = s_j$, leaving the endpoint state unchanged. □

Proposition 2 (Termination of rewriting and shortest-path minimality in the observed graph)**.** *Rules (R1)–(Rℓ) strictly decrease word length, so rewriting terminates in $\leq k$ steps at a sound, simple observed path. The BFS word W' is no longer than any path obtainable by such loop deletions; in particular, $|W'| \leq |W|$.*

Proof. Each rule removes ≥ 1 symbol, and length is a non-negative integer, giving termination; each deletion preserves the endpoint by Proposition 1. A fixed point has no repeated state, since any repetition exposes an (Rk)-loop. Every remaining edge is in G_W, so every rewriting result is an s_0–s_k path in G_W. BFS is shortest among all such paths and therefore is no longer than any rewriting result or the original walk. □

Removing (Rℓ)-loops can only shorten and cannot break the solution (Proposition 1); the residual gap to the true geodesic is the non-commutative-detour term (ii), which requires knowledge of transitions the teacher never took. Our experiments quantify this gap and show it is small (Section 6). ATR runs in $O(k)$ time and space per trajectory.

Empirical vs. global geodesic. We distinguish three objects the reviewer-minded reader should keep separate: the *global geodesic* $d(s_{\text{init}}, s_{\text{goal}})$ (shortest word over *all* of $\mathcal{M}$), the *empirical geodesic* W' (shortest path in the observed graph G_W), and a *rewriting fixed point* (a walk with no repeated state). Proposition 2 shows that BFS is at least as short as the latter; they need not coincide because merging all repeated occurrences in G_W can expose shortcuts that no particular sequence of loop deletions selects. The first is a lower bound, $|W'| \geq d(s_{\text{init}}, s_{\text{goal}})$. The following makes the boundary precise.

Proposition 3 (When ATR recovers the true geodesic)**.** *(a) If G_W contains some globally shortest solving path—i.e. the teacher's transitions* cover *a geodesic—then* ATR *returns a true geodesic, $|W'| = d(s_{\text{init}}, s_{\text{goal}})$. (b) If the teacher's redundancy consists only of inserted state-relative identity sub-words around an underlying geodesic (no term-(ii) detour), then* ATR *equals the oracle: $|W'| = d(s_{\text{init}}, s_{\text{goal}})$. (c) In general $d(s_{\text{init}}, s_{\text{goal}}) \leq |W'| \leq |W|$.*

Proof. (a) BFS returns a shortest path in G_W; if a global geodesic is a subgraph of G_W then no shorter path exists, so W' attains d. (b) Inserting loops preserves every edge of the underlying geodesic in G_W, so the claim follows from (a), independently of rewriting order. (c) The lower bound is definitional and the upper bound follows from Proposition 2. The empirical gap is exactly $\text{dist}_{G_W}(s_{\text{init}}, s_{\text{goal}}) - d(s_{\text{init}}, s_{\text{goal}})$. □

Proposition 3(a) is a *coverage* condition: ATR is optimal exactly when the teacher (or a pool of teachers whose graphs are unioned) has *seen* a geodesic's transitions, even if no single trajectory took it. This is why pooling many trajectories into one G_W before extraction can recover shortcuts no individual trajectory contains, and it delimits precisely when search (Zhou et al., 2024a) or preference optimization (Putta et al., 2024) is needed instead (term (ii) with no covering observation).

When ATR coincides with, or is sounder than, a dedup baseline. A natural engineering baseline removes failed and duplicate steps ("DEDUP": drop any generator whose response is an error, empty, or a repeat of an earlier one). The next proposition gives a sufficient collapse condition and a separating case.

Proposition 4 (Monotone collapse and revisit separation)**.** *Consider a response-integration state I in which each realized call either leaves I unchanged (failure, empty, or duplicate response) or maps it to $I \cup \{r\}$ for a new response r. (a) Applied to a single trajectory in this strictly monotone quotient,* ATR *equals* DEDUP*. (b) Whenever the observed state sequence may revisit a state, as in sequential tool-use,* DEDUP *can be* unsound *—removing a duplicated observation may change the endpoint state—whereas* ATR*, taking a shortest path in the observed state graph, is sound under deterministic tools and exact state identifiers (Proposition 1) and never longer than the input.*

Proof. (a) A new response strictly enlarges I, so informative states cannot repeat; failures, empty responses, and duplicates are precisely self-loops. After removing self-loops, the single observed walk is a directed chain, whose unique BFS path keeps each new response once in observed order—the DEDUP output. (b) Consider the walk $\texttt{go}_A\,\texttt{go}_B\,\texttt{go}_A$ where the two $\texttt{go}_A$ steps re-observe location A identically (so DEDUP treats the second as a duplicate) while the agent's location genuinely changes: DEDUP deletes the second $\texttt{go}_A$ and returns $\texttt{go}_A\texttt{go}_B$, ending at $B \neq A$; ATR finds the state revisit $A \to B \to A$ and returns $\texttt{go}_A$, ending at A. □

Proposition 4 clarifies the empirical picture in Section 5: on TOOLBENCH, where realized responses form this information-integration quotient, ATR collapses to the strongest hand-designed baseline by construction, while on sequential/embodied tool-use (ALFWorld, web agents such as WEBSHOP and WEBARENA (Yao et al., 2022; Zhou et al., 2024b)) the monoid is non-commutative and ATR can preserve reachability where a dedup baseline fails.

Soundness under partial observability. When state identifiers are partial (two different global states share an observation), merging them can be unsound. Our domain-specific certificate (Algorithm 2) assumes that equal observations identify every component alterable by navigation/inspection, and marks every action that can alter any remaining hidden component as mutating. It collapses $o_i = o_j$ only if every intervening action is non-mutating. *Positive example*

Algorithm 2 Certified merge under partial observability

Require: observations $o_0, \ldots, o_k$, actions $t_1, \ldots, t_k$, a complete hidden-state mutation predicate μ under the domain observation model

1: **for** earliest i with $o_i = o_j$ for some $j > i$; take the largest such j **do**
2: **if** $\mu(t_{i+1}) = \cdots = \mu(t_j) =$ FALSE **then**
3: delete steps $i{+}1, \ldots, j$; **restart**
4: **end if**
5: **end for**
6: **return** remaining walk

(ALFWorld): `go to shelf 1`; `examine shelf 1`; `go to shelf 1` returns the same observation with only navigation/inspection in between—safely collapsed. *Negative example*: `open fridge`; `take apple`; `close fridge` may re-observe "`The fridge is closed`" identical to before opening, yet the inventory changed—the intervening `take` is mutating, so the rule refuses the merge. The cost of this conservatism is coverage: on real GPT-4 ALFWorld data it certifies detours in 9.2% of trajectories (Section 5); a richer state abstraction (tracking inventory/pose) would certify more.

5 Redundancy and Non-Commutativity in Real Agent Data

We study two real regimes that instantiate the two cases of Proposition 4: information-monotone response integration (ToolBench) and sequential embodied tool-use (ALFWorld). Table 1 and Fig. 1 summarize the comparison; we begin with the embodied case, then the API case. We first analyze the ALFWorld split of AgentInstruct (Zeng et al., 2024), 336 GPT-4 trajectories ($4{,}516$ actions, mean length 13.4) in a text-embodied household (Shridhar et al., 2021). Observations serve as (partial) state identifiers.

The action monoid is non-Abelian. Preconditions are pervasive and order-dependent: 100% of the 484 successful `take` actions are enabled by a prior `go to` their receptacle, and 93.7% of the 412 `put` actions occur while holding the object. Equivalently, reversing these orders yields the environment's no-op response—empirical confirmation that $t_a t_b \neq t_b t_a$. Only 0.04% of actions carry the explicit `"Nothing happens."` failure marker; this is narrower than either repeated-observation no-ops or the simulator admissibility metric used in closed-loop evaluation.

Sound detours exist. Using the certified navigation-only merge (Section 4), 9.2% of trajectories contain a provably state-preserving detour; ATR removes 57 actions in total (1.3%), e.g. an agent that examines `sidetable 1` and walks back with no state change. Curated GPT-4 data is thus nearly geodesic at the action-validity level—an honest finding—while the algebraic *structure* ATR exploits (non-commutativity, identity loops) is unmistakably present. *(To pre-empt confusion:* this conservative 1.3% is what is certified under the stated ALFWorld mutation model; the larger 24.4% in Table 1 uses the more aggressive observation-as-state proxy. Its 1.5% offline replay-failure floor is already present in Raw because the cached observation/action map is ambiguous; these are lower and upper reduction operating points, not contradictory rates.*)* Redundancy grows with teacher verbosity and with search-based collection; we study this dependence directly in a controlled setting next, where teacher quality is a knob rather than a constant.

Where redundancy is large: search-collected data. Curated single-answer data is near-geodesic, but the trajectories that dominate modern agent tuning are *search-collected* and self-sampled. A canonical case is ToolBench (Qin et al., 2024), whose supervised data—the data ToolLLaMA is trained on—is produced by ChatGPT/GPT-4 exploring a *depth-first search tree* (DFSDT) over 16k real APIs; the retained answer paths therefore still carry the debris of search. On the full public DFSDT train set ($44{,}730$ successful trajectories, $142{,}736$ calls) we measure a $\mathbf{27.9}\%$ *failed-call* rate—each a no-op in the response-integration state, an errored API call that leaves the information state unchanged—a 4.2% empty-response rate, and a $\mathbf{14.3}\%$ *query-construction* dependency rate between successive calls (a later call's arguments are taken from a predecessor's response). We stress that this is a dependency in how queries are *built*; it is a dependency of call generation. Our

Table 1: Reduction and offline observed-transition replay on real data. On the ToolBench response-integration quotient, ATR coincides with B2; on sequential ALFWorld the observation-proxy ATR removes twice as much and cuts replay failures from 25.3% to the 1.5% ambiguity floor already present in RAW. The certified variant removes 1.3% under its stated mutation assumptions. Reduction is % removed vs. RAW.

Dataset	Regime	B1	B2	B3	ATR	Replay fail. R/B2/A
ToolBench (44,730 traj.)	info.-monotone	27.9	**39.4**	32.9	**39.4**	n/a
ALFWorld (336 traj.)	sequential	8.6	12.0	10.9	**24.4**	$1.5/25.3/1.5\%$

reduction instead acts on the quotient that integrates already-realized responses into a *monotonically accumulating* set. Those union updates commute, although the grounded calls that produced them need not. ATR's information-state reduction removes $\mathbf{39.4}\%$ of all calls (mean calls per task $3.19 \rightarrow 1.93$).

A by-construction coincidence, verified. ToolBench realized-response updates instantiate Proposition 4(a), so ATR reduces by construction to B2: both reach 1.93 mean calls, while error-only B1 reaches 2.30 and adjacent-response B3 2.14. The empirical finding is that failed/empty/duplicate responses account for 39.4% of calls in this search-collected corpus; equality with B2 is an implementation check, not evidence separating methods. ToolBench therefore cannot separate them in closed-loop, so we do not use it for our distillation study. The place where ATR can be sounder than DEDUP is state-revisiting, sequential tool-use (Proposition 4(b)); we therefore run real closed-loop distillation there using the same 336 curated GPT-4 demonstrations and use the controlled Delivery Workshop (Section 6) to isolate the mechanism causally. The accompanying code ships this ALFWorld track (RAW vs. B2 vs. ATR SFT of a QWEN2.5 student, scored by the free ALFWorld simulator with no external judge), plus an optional STABLETOOLBENCH pipeline for the ToolBench regime.

Sequential real data: ATR $\neq$ DEDUP, and DEDUP is unsound. Proposition 4(b) predicts that state revisits can make a dedup baseline unsound. We observe this separation on the 336 AgentInstruct ALFWorld trajectories, using the observation as a state proxy. ATR and DEDUP produce *different* reductions on $\mathbf{25.3}\%$ of trajectories (vs. never on ToolBench); ATR removes $\mathbf{24.4}\%$ of actions against DEDUP's 12.0%—twice as much. In offline replay against the cached $(\text{observation}, \text{action})$ transition map, DEDUP's output fails to reach the original observed endpoint on $\mathbf{85}/\mathbf{336}$ (25.3%) of trajectories—the same 85 on which B2 and ATR differ. Of these, 71 encounter an unobserved state–action pair and 14 reach a different endpoint, because deleting a step that merely re-observes a location silently drops the state change that a later step depended on. Both RAW and ATR replay on $331/336$; their shared five failures expose ambiguity already present in the cached observation/action map, so ATR adds no failures over the RAW floor. All 336 trajectories are retained for SFT. Thus the aggressive proxy is more effective and empirically sounder than B2, without inheriting the exact-state guarantee of Proposition 1.

Closed-loop distillation: compressed supervision preserves performance. We SFT QWEN2.5-1.5B and 3B students (LoRA, action-only assistant targets, identical optimizer budgets) on RAW/B2/ATR supervision and evaluate all models on the same 134 unseen ALFWorld games over 3 seeds (Table 2). At 1.5B, success is at floor, but ATR reduces inadmissible actions and no-ops; a repeat of the identical ATR seed-2 adapter obtains $1/134$ successes, confirming that the reported zero mean is not a systematic inability to complete tasks (Section F). At 3B, success rises to 9–10% and the three supervisions remain within seed variation: ATR reaches $9.5_{\pm 1.9}\%$ success versus RAW's $9.2_{\pm 0.3}\%$ and B2's $10.0_{\pm 1.5}\%$, while improving no-op rate over B2 (3.7% vs. 5.1%). Thus the 3B result supports *no detectable degradation within this three-seed evaluation*, not a claim of universal ATR superiority: ATR removes 24.4% of supervised action turns yet remains comparable to the uncompressed teacher. Low absolute success and the small seed count cannot exclude subtle downstream differences; in particular, B2's replay failures do not yield a detectable 3B penalty, plausibly because SFT need not reproduce any demonstration verbatim and can learn recovery actions across trajectories. Shorter trajectories offer corresponding token/FLOP savings under token-budgeted or

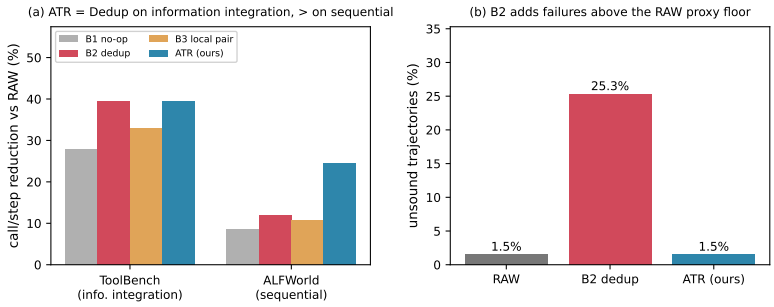

Figure 1: Real-data comparison across the cases in Proposition 4. (a) ATR equals B2 on the information-monotone ToolBench quotient but removes $2\times$ more on sequential ALFWorld. (b) Offline replay fails for 25.3% of B2 paths versus 1.5% of observation-proxy ATR paths, equal to the RAW ambiguity floor.

Table 2: Closed-loop ALFWorld distillation on 134 unseen games (3 seeds; simulator-scored). ATR uses 24.4% fewer action targets than RAW (10.16 vs. 13.44 steps/trajectory). At 1.5B, it improves action validity while success is at floor; at 3B, all supervisions are similar within seed variation, so ATR preserves performance despite shorter supervision. Mean$\pm$std in percentage points.

Student	Supervision	Success↑	GoalCond.↑	Inadm.-act.↓	No-op↓
1.5B	RAW	$1.7_{\pm 0.4}$	$38.4_{\pm 3.1}$	$39.8_{\pm 3.4}$	$20.5_{\pm 4.5}$
	B2	$0.5_{\pm 0.7}$	$36.9_{\pm 2.6}$	$35.3_{\pm 1.9}$	$19.6_{\pm 1.7}$
	ATR	$0.0_{\pm 0.0}$	$33.6_{\pm 5.2}$	$\mathbf{30.7_{\pm 4.4}}$	$\mathbf{13.4_{\pm 4.0}}$
3B	RAW	$9.2_{\pm 0.3}$	$58.5_{\pm 2.0}$	$18.9_{\pm 3.8}$	$3.4_{\pm 1.7}$
	B2	$10.0_{\pm 1.5}$	$56.3_{\pm 1.4}$	$21.6_{\pm 1.7}$	$5.1_{\pm 1.5}$
	ATR	$9.5_{\pm 1.9}$	$56.3_{\pm 1.4}$	$21.3_{\pm 1.9}$	$3.7_{\pm 1.0}$

dynamically padded SFT; we do not claim an equal wall-clock reduction because observation tokens and padding also contribute.

ALFWorld metrics (Table 2). *Success* is the simulator terminal reward; *GoalCond.* the fraction of subgoals satisfied; *Inadm.-act.* and *No-op* the per-step fractions of actions outside the admissible set and leaving the observation unchanged; and generated tokens/step a decoding-conciseness diagnostic.

6 Controlled Study: The Delivery Workshop

Environment. We instantiate the theory in a transformation-monoid environment with full control and computable geodesics. The state is $(\text{agent location}, \text{hand}, \text{item positions})$ over $L{=}6$ locations and $K{=}4$ items; generators are $\texttt{go}_\ell$, $\texttt{take}_i$, $\texttt{put}_i$, and `look` (15 tools total). `take` requires co-location and an empty hand; `put` requires holding—so the monoid is non-Abelian with the same precondition structure measured in Section 5. A task delivers a random subset of items to random targets. We sample 6,000 train and 1,000 eval tasks.

Teacher and supervision regimes. A *verbose teacher* produces *correct* trajectories that inject only provably-removable redundancy—`look` no-ops, backtracking loops (`go a;go back`), and `take;put` cancellations— plus a randomly ordered (possibly sub-optimal) delivery plan, mirroring documented weak/strong failure modes (Shinn et al., 2023; Liu et al., 2024). From each teacher trajectory we derive three supervisions: RAW (as-is), ATR (Algorithm 1), and ORACLE (BFS shortest word; requires the full model, a skyline). Averaged over the train set, mean lengths are RAW 17.7, ATR 9.6, ORACLE 9.1: ATR removes 45.8% of steps and lands within 0.47 of the true geodesic (Fig. 3), i.e. the model-free reducer nearly matches optimal planning.

Table 3: Behavioral cloning under three supervisions (5 seeds, 6,000 train tasks, identical model and update budget). ATR matches the model-free-infeasible ORACLE skyline on success while nearly eliminating non-commutativity errors, using 46% fewer supervision transitions than RAW.

Supervision	Train trans.	Success ↑	Invalid-act. ↓	Exact-opt. ↑	Opt. gap ↓
RAW (verbose teacher)	106,271	$0.714_{\pm.016}$	$0.409_{\pm.014}$	$0.565_{\pm.015}$	$0.265_{\pm.016}$
ATR (ours)	**57,584**	$\mathbf{0.972}_{\pm.003}$	$\mathbf{0.007}_{\pm.003}$	$0.731_{\pm.011}$	$0.313_{\pm.018}$
ORACLE (skyline)	54,785	$0.978_{\pm.004}$	$0.051_{\pm.012}$	$0.927_{\pm.008}$	$0.056_{\pm.007}$

Policy, training, evaluation. The student is a state-conditioned policy $\pi(a \mid \text{goal}, s)$—a 2×256 MLP over a 67-d structured feature vector. Architecture, optimizer, and the *total number of gradient updates* (3,000) are held identical across regimes. This compute-matched design changes only the supervision pool, but shorter pools are revisited more often; an equal-epoch control in Section C removes this possible confound. We evaluate *closed-loop*: roll the policy out in the environment (invalid actions are no-ops) for up to 40 steps and report task success, executed length, optimality gap vs. the oracle geodesic, exact-optimal rate, and the invalid-action (non-commutativity error) rate. All numbers are mean±std over 5 seeds. *Invalid-action rate* is the per-step fraction of precondition-violating (no-op) actions; *optimality gap* and *exact-optimal* are computed on successful episodes only (normalized excess length over the oracle geodesic, and the fraction with length $= \ell^\star$), so they are success-conditioned and not comparable across rows of differing success (full definitions in Section C). Invalid-action rate is unconditional; ORACLE's higher rate (0.051 vs. ATR's 0.007) is consistent with its narrower shortest-path supervision covering fewer recovery states.

Main result. Table 3 shows that cloning ATR-reduced data raises success from 0.714 to 0.972—statistically indistinguishable from the ORACLE skyline (0.978)—while *using* 46% *fewer supervision transitions*. The mechanism is visible in the invalid-action rate: the RAW student, having been taught to emit no-ops and backtracks, produces an invalid action 40.9% of the time and exhausts its budget; ATR collapses this to 0.7%. ATR trails ORACLE only on *exact*-optimality (0.731 vs. 0.927): the residual gap is the non-commutative ordering term (ii), which ATR cannot fix from observed transitions alone—consistent with Proposition 3. The apparently lower RAW optimality gap is survivorship bias: this success-conditioned metric excludes the many harder tasks that RAW fails.

Data efficiency and robustness. ATR tracks the ORACLE skyline at every training budget (at 1,000 tasks: 0.79 vs. RAW 0.16). As teacher redundancy increases, RAW success falls from 0.93 to 0.24, whereas ATR stays near 0.93 with constant supervision length. The full curves and distributions are in Figs. 3 and 4; removing algebraic noise both shortens sequences and makes each task more informative.

7 CONCLUSION AND LIMITATIONS

Viewing tool use as a transformation-monoid word defines useful supervision as a geodesic, free of state-relative identity loops and ordering detours. ATR extracts a sound, model-free empirical geodesic in linear time under exact state identifiers. On real GPT-4 data, observation-proxy reduction shortens supervision with no detectable 3B degradation; in the controlled full-state study, it closes nearly all of the naive-cloning-to-oracle gap.

ATR deliberately removes only redundancy certifiable from observed states. Ordering errors requiring unseen transitions instead need search (Zhou et al., 2024a) or preference optimization (Putta et al., 2024; Rafailov et al., 2023); gains over dedup are therefore regime-dependent. The 24.4% ALFWorld operating point uses an observation proxy with 1.5% offline replay failures, whereas the certified variant removes only 1.3%; defining mutation predicates is harder in open-ended environments. Reported SFT is action-only, and extending ATR to supervised reasoning traces requires consistency handling. Finally, low Qwen success and three seeds make that experiment an initial scaling test, not evidence of universal superiority or strict statistical equivalence.

AI USE STATEMENT

Generative AI tools assisted research ideation, related-work discovery, drafting and language editing, code and debugging, and checking mathematical arguments. The authors independently verified all claims, citations, implementations, experimental protocols, and results, and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

All environments, the ATR algorithm, training, evaluation, ablations, and figure generation are implemented in the accompanying code; the real-data analyses download public datasets (TOOLBENCH, AGENTINSTRUCT). Algorithms 1 and 2 and Propositions 1 to 4 fully specify the method; Section 6 lists the environment (L=6, K=4), model (2×256 MLP, 67-d features), update budget (3,000), evaluation budget (40), and seed count (5). The real-data reductions and the Table 1/Fig. 1 numbers are produced by scripted CPU pipelines; the primary closed-loop distillation track (ALFWorld, QWEN2.5 students, simulator-scored) and an optional STABLETOOLBENCH track ship as turnkey scripts with offline model/data staging. Every reported number is produced by a single scripted pipeline.

REFERENCES

Franz Baader and Tobias Nipkow. *Term Rewriting and All That*. Cambridge University Press, 1998.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *AAAI Conference on Artificial Intelligence*, 2024.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-FLAN: Designing data and methods of effective agent tuning for large language models. *Findings of the Association for Computational Linguistics (ACL)*, 2024.

David B A Epstein, James W Cannon, Derek F Holt, Silvio V F Levy, Michael S Paterson, and William P Thurston. *Word Processing in Groups*. Jones and Bartlett, 1992.

Richard E Fikes and Nils J Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3-4):189–208, 1971.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

John M Howie. *Fundamentals of Semigroup Theory*. Oxford University Press, 1995.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced reasoning and learning for autonomous AI agents. *arXiv preprint arXiv:2408.07199*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *International Conference on Learning Representations (ICLR)*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research (TMLR)*, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. Lumos: Learning agents with unified data, modular design, and open-source LLMs. *Findings of the Association for Computational Linguistics (ACL)*, 2024.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. *Findings of the Association for Computational Linguistics (ACL)*, 2024.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning (ICML)*, 2024a.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024b.

A ATR ALGORITHM DETAILS

Deterministic rewriting. The local rewriting view repeatedly deletes the longest sub-walk anchored at the earliest revisited state:

1: **input** states $s_0, \ldots, s_k$, actions $a_1, \ldots, a_k$
2: **repeat**
3: find the earliest i with $s_i = s_j$ for some $j > i$; take the largest such j
4: delete states $s_{i+1..j}$ and actions $a_{i+1..j}$
5: **until** no repeated state
6: **return** the remaining walk

Each iteration strictly shortens the walk, so it terminates in $\leq k$ steps at a sound simple path. The BFS empirical geodesic can be strictly shorter because the merged observed graph may expose additional combinations of observed edges (Proposition 2); it is the operational ATR output used in every experiment. BFS runs in $O(k)$ time and space over at most $k+1$ vertices and k edges.

Pooled extraction. Proposition 3(a) motivates unioning the per-trajectory graphs $G_{W_1} \cup \cdots \cup G_{W_m}$ over trajectories that share $(s_{\text{init}}, s_{\text{goal}})$ before BFS; a geodesic in the union may use edges from several teachers, recovering shortcuts no single trajectory contains. All reported ATR numbers use the conservative per-trajectory version unless stated otherwise.

B REAL-DATA PROCESSING

B.1 TOOLBENCH (INFORMATION-MONOTONE RESPONSE INTEGRATION)

Source. The public `toolllama_G123_dfs_train` file (Qin et al., 2024): 180k DFSDT path objects; we keep the 44,730 that end in `Finish(give_answer)` with $\geq$ 2 calls. Each turn is `system`/`user`/`assistant`/`function`; an assistant turn is `Thought:` ... `Action:` a `Action Input:` $json$ and a function turn is {`"error":`e,`"response":`r}.

No-information classifiers. A realized call is a self-loop in the response-integration state if its response (i) has non-empty `error` or matches a failure marker (rate limit, does not exist, invalid key, timeout, ...), (ii) is empty (`num_results:0`, empty list), or (iii) byte-duplicates an earlier kept response. DEDUP(B2) removes (i)–(iii); B1 removes only (i); on ToolBench, B3 removes (i) and collapses each pair of adjacent byte-identical responses to one; ATR takes the info-state geodesic (equal to B2 here by Proposition 4a); ORACLE-sub additionally drops informative calls unreferenced in the final answer (a heuristic length skyline).

Query-construction dependency evidence. A successive call is *query-dependent* if a token in its `Action Input` appears in an earlier response body. We first remove errored calls and count successive pairs within each trajectory, giving 61,976 pairs, of which 14.3% meet the criterion. This call-construction dependency is outside the response-integration quotient: retained responses still accumulate by order-independent union, so ATR=B2 here (Proposition 4a). Per-condition mean calls: RAW 3.19, B1 2.30, B3 2.14, B2 = ATR 1.93, ORACLE-sub 1.50.

B.2 ALFWORLD (SEQUENTIAL / NON-COMMUTATIVE REGIME)

Source. AGENTINSTRUCT ALFWorld split (Zeng et al., 2024), 336 GPT-4 trajectories. The state proxy is the observation string before each action; a `"Nothing happens."` response is a certified no-op (state unchanged). Episode-control meta-actions (`Task succeeded.`/`End of task`) are stripped before analysis.

Reductions. B1 drops no-ops; DEDUP(B2) drops any action whose resulting observation duplicates an earlier one; B3 drops adjacent `open`/`close` inverse pairs; ATR is the BFS empirical geodesic over observed transitions. We replay offline using the first cached (observation, action) $\mapsto$ next observation transition; a missing edge or different final observation is a replay failure, not an interactive simulator rerun. Results: mean length RAW 13.44, B1 12.28, B3 11.98, B2 11.83, ATR 10.16 (ATR -24.4% vs. B2 -12.0%); ATR and B2 differ on 25.3% of trajectories, exactly the

$85/336$ B2 replay failures (71 missing edges, 14 wrong endpoints). RAW and ATR each fail on the same $5/336$ because the cached proxy itself is ambiguous; thus ATR adds no failures over RAW. All 336 examples are retained in the SFT sets.

C CONTROLLED STUDY: FULL SPECIFICATION

Environment. States $(loc, hand, pos)$ with $L{=}6$ locations, $K{=}4$ items; generators $\texttt{go}_\ell$ (6), $\texttt{take}_i$ (4), $\texttt{put}_i$ (4), `look`. $\texttt{take}_i$ requires $hand{=}\varnothing \wedge pos_i{=}loc$; $\texttt{put}_i$ requires $hand{=}i$; otherwise the generator acts as identity. Tasks deliver a random item subset to random targets (6,000 train / 1,000 eval). The oracle geodesic is BFS over the reachable state space; ATR is the empirical geodesic of the teacher walk.

Verbose teacher. Correct trajectories with injected state-relative identity redundancy: `look` no-ops (prob. 0.35), backtracking loops `go a; go back` (prob. 0.40), and `take; put` cancellations (prob. 0.25), plus a random delivery order. The redundancy sweep scales these to $0/0.5/1.0/1.5/2.0\times$ (Table 5).

Policy & training. Featurizer: 67-d binary vector (agent loc one-hot, hand one-hot, per-item position and goal one-hots). Policy: MLP $67{\to}256{\to}256{\to}15$, ReLU. Training: 3,000 AdamW updates, batch 1024, lr $3{\times}10^{-3}$ cosine, weight decay 10^{-4}, identical across regimes; evaluation is a closed-loop rollout, budget 40; 5 seeds.

Equal-epoch control. The main comparison matches compute at 3,000 updates, so shorter pools receive more expected presentations per transition. We additionally match RAW's 28.9 expected epochs: RAW uses 3,000 updates, ATR 1,626, and ORACLE 1,547 (batch 1024, 5 seeds). Success is $0.714_{\pm.016}/0.982_{\pm.004}/0.979_{\pm.003}$ and invalid-action rate is $0.409_{\pm.014}/0.003_{\pm.003}/0.046_{\pm.012}$ for RAW/ATR/ORACLE, respectively. Thus unequal sample re-exposure does not explain ATR's gain.

Metric definitions (Table 3). *Invalid-action rate*: the per-step fraction of emitted actions whose precondition fails (a no-op leaving the state unchanged), averaged over all executed steps of all episodes. *Optimality gap*: computed on *successful* episodes only, as the normalized excess length $(\text{executed length} - \ell^\star)/\ell^\star$, with $\ell^\star$ the oracle-geodesic length for the task; failed episodes are excluded (no well-defined completion length), so the gap is success-conditioned and *not* comparable across rows with different success rates; in particular, RAW's apparently smaller gap reflects survivorship on easier solved tasks. *Exact-optimal rate*: the fraction of successful episodes with executed length $= \ell^\star$. This unconditional invalid-action rate is higher for ORACLE than ATR (0.051 vs. 0.007), consistent with shortest-word supervision visiting fewer recovery states; both remain far below RAW (0.409).

Table 4: Data-efficiency (success rate vs. # training tasks, 3 seeds).

# tasks	250	500	1000	2000	4000	6000
RAW	0.075	0.111	0.164	0.332	0.570	0.710
ATR	0.259	0.534	0.790	0.906	0.953	0.973
ORACLE	0.241	0.532	0.782	0.921	0.967	0.977

Table 5: Teacher-redundancy sweep (3,000 train tasks, 1,000 eval tasks, 3,000 updates, 3 seeds): mean supervision length and success.

level	$0.0\times$	$0.5\times$	$1.0\times$	$1.5\times$	$2.0\times$
RAW length	9.68	13.49	17.99	21.70	25.60
ATR length	9.68	9.70	9.77	9.63	9.64
RAW success	0.931	0.655	0.453	0.335	0.242
ATR success	0.931	0.939	0.938	0.932	0.935

D ADDITIONAL CONTROLLED-STUDY RESULTS

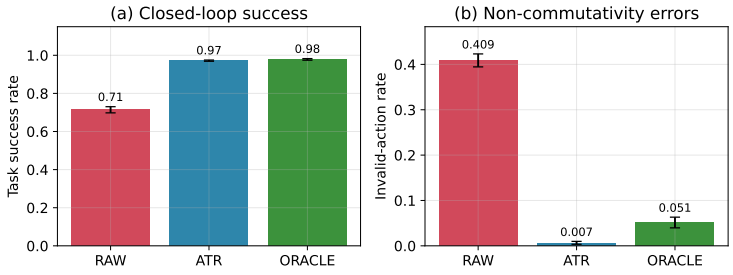

Figure 2: Closed-loop evaluation (5 seeds). (a) ATR matches the ORACLE skyline and far exceeds RAW. (b) ATR nearly eliminates precondition (non-commutativity) errors.

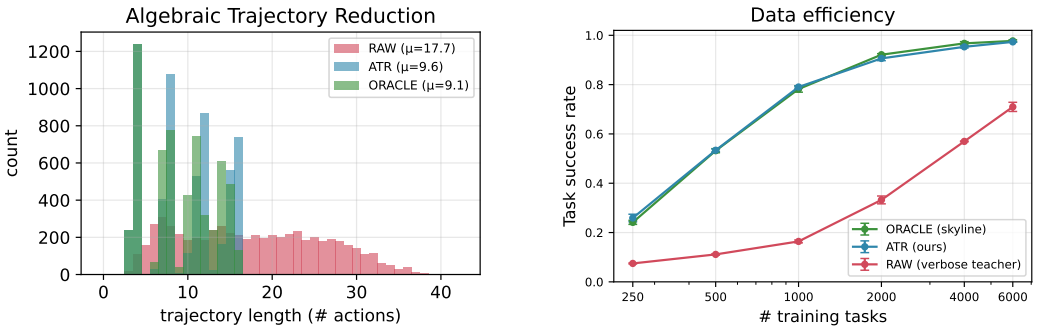

Figure 3: *Left:* supervision-length distribution; ATR compresses RAW by 45.8% to within 0.47 steps of ORACLE. *Right:* data efficiency: ATR tracks the ORACLE skyline at every budget while RAW lags badly.

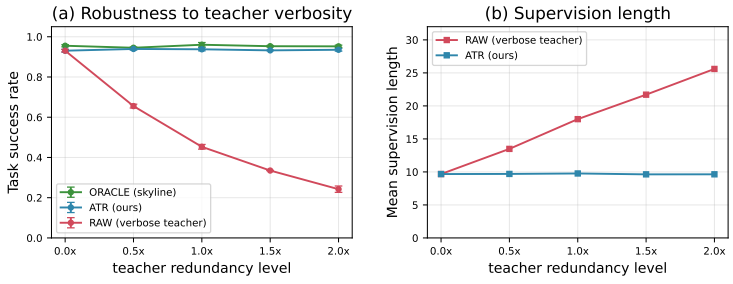

Figure 4: Teacher-redundancy sweep. (a) RAW degrades monotonically as the teacher gets more verbose; ATR stays flat and near ORACLE. (b) ATR supervision length is constant while RAW grows linearly.

E FULL ALFWORLD STUDENT DIAGNOSTICS

Table 6: Closed-loop ALFWorld distillation (Qwen2.5-1.5B LoRA, 3 seeds, 134 unseen games; the simulator scores every metric, no external judge). The 1.5B student is too weak to solve ALFWorld, so task *success* and *goal-condition* sit near the floor and are statistically tied across supervisions (ATR's 0.0 is floor noise, not a systematic failure: an identical-adapter rerun scores $1/134$, Section F). The mechanism from the controlled study nevertheless transfers: cloning ATR-reduced trajectories yields the *cleanest* policy—the lowest inadmissible-action and no-op rates (RAW $>$ B2 $>$ ATR), confirming that ATR removes the provably-redundant residue the student would otherwise imitate. Metric definitions in Section F; bold marks the best mean per column.

Supervision	Success↑	GoalCond.↑	Inadm.-act.↓	No-op↓	Gen. tok/step
RAW (verbose teacher)	$\mathbf{1.7}_{\pm \mathbf{0.4}}$	$\mathbf{38.4}_{\pm \mathbf{3.1}}$	$39.8_{\pm 3.4}$	$20.5_{\pm 4.5}$	$18.89_{\pm 0.62}$
B2 dedup	$0.5_{\pm 0.7}$	$36.9_{\pm 2.6}$	$35.3_{\pm 1.9}$	$19.6_{\pm 1.7}$	$19.50_{\pm 0.60}$
ATR (ours)	$0.0_{\pm 0.0}$	$33.6_{\pm 5.2}$	$\mathbf{30.7}_{\pm \mathbf{4.4}}$	$\mathbf{13.4}_{\pm \mathbf{4.0}}$	$19.28_{\pm 0.67}$

Table 7: Closed-loop ALFWorld distillation (Qwen2.5-3B LoRA, 3 seeds, 134 unseen games; the simulator scores every metric, no external judge). Same supervision (RAW/B2/ATR) and optimizer budget across conditions; metric definitions in Section F; bold marks the best mean per column.

Supervision	Success↑	GoalCond.↑	Inadm.-act.↓	No-op↓	Gen. tok/step
RAW (verbose teacher)	$9.2_{\pm 0.3}$	$\mathbf{58.5}_{\pm \mathbf{2.0}}$	$\mathbf{18.9}_{\pm \mathbf{3.8}}$	$\mathbf{3.4}_{\pm \mathbf{1.7}}$	$23.08_{\pm 0.25}$
B2 dedup	$\mathbf{10.0}_{\pm \mathbf{1.5}}$	$56.3_{\pm 1.4}$	$21.6_{\pm 1.7}$	$5.1_{\pm 1.5}$	$21.71_{\pm 0.69}$
ATR (ours)	$9.5_{\pm 1.9}$	$56.3_{\pm 1.4}$	$21.3_{\pm 1.9}$	$3.7_{\pm 1.0}$	$22.75_{\pm 0.20}$

F CLOSED-LOOP DISTILLATION PROTOCOL (REPRODUCIBLE)

Our closed-loop experiments distil QWEN2.5-1.5B and 3B students with LoRA (r=16, α=32, dropout 0.05; all attention+MLP projections), AdamW at 10^{-4}, a cosine scheduler configured for 1,500 steps with 3% warmup, micro-batch 2, gradient accumulation 8, 1,500 micro-batches (187 optimizer updates), maximum length 2,048, gradient clipping 1.0, and bf16 where supported. Loss is applied only to assistant tokens, with an identical budget across conditions; offline model/data staging for network-restricted servers. Table 2 reports both scales; Tables 6 and 7 give all diagnostics separately. **(i) ALFWorld (primary, sequential regime):** SFT on RAW/B2/ATR supervision from the same 336 curated GPT-4 trajectories (where ATR$\neq$B2 already; Table 1). These converted ALFWorld targets contain only `ACTION: ...`, not teacher thoughts, and retain all 336 trajectories, including the five with proxy replay ambiguity. The released pipeline also supports weak-teacher self-sampling, but those trajectories are not used in the reported student results. Closed-loop success, goal-condition score, inadmissible-action and no-op rates, and generated tokens are scored by the simulator—*no external judge*. This is the regime where the data analysis finds a reduction advantage over B2; student behavior is reported separately rather than implied by Proposition 4(b). *ALFWorld metric definitions (Table 2): Success* = task solved (terminal simulator reward); *GoalCond.* = fraction of the task's goal conditions satisfied (graded partial credit, parsed from the game's subgoal decomposition, so it grades progress even on unsolved episodes); *Inadm.-act.* = per-step fraction of emitted actions *not* in the simulator's admissible-command set (the ground-truth invalid rate); *No-op* = per-step fraction of actions that leave the observation unchanged; *Gen. tok/step* = mean tokens generated per action at inference (a conciseness diagnostic, tied across conditions; *not* a training-token count). *Reproducibility of the eval.* Greedy decoding is used, but GPU floating-point non-determinism can occasionally flip a near-tied argmax, so metrics are not bit-identical across reruns. In a control we re-ran ATR seed 2 on the identical adapter: the mechanism metrics are stable to within noise (inadmissible-action $0.354 \to 0.359$, no-op $0.190 \to 0.189$, goal-condition $0.287 \to 0.289$), while task success fluctuates at the floor ($0/134 \to 1/134$). We report the first run for every seed without selection; the 3-seed standard deviations in Table 2 already exceed this run-to-run jitter. Optional scripts also support STABLETOOLBENCH; because ATR$\equiv$B2 in its response-integration quotient, this is not a central reported comparison.