

Certified Decision-Equivalent Context Compression for LLM Agents

Chandra Shekhar Mudarapu*

July 17, 2026

Abstract

LLM agents resend a growing context every turn, so context size dominates serving cost—yet every shipping compressor quotes token savings with *no guarantee the agent still behaves the same*. We reframe the objective from byte- or embedding-fidelity to **decision-equivalence**: a compression is acceptable iff the agent takes the same next action it would have on the uncompressed context. We make this objective *certifiable* with a distribution-free, finite-sample guarantee via conformal risk control (Learn-Then-Test / CRC), the per-turn loss being a decision flip, and validate it out-of-sample on real SWE-bench traces (coverage 96.6%–100% at the 95% target). Certifying a per-turn proxy is necessary but not sufficient: on a real end-to-end agent (SWE-bench Verified, official harness) we show—without hedging—that *aggressive lossy* compression does not transfer to task success. The remedy is architectural: a **reversible, relevance-gated** engine that digests only aged-out periphery while keeping the working set intact and recoverable. On a 500-instance, 30-turn long-horizon agent it is the **highest-accuracy compressor** (36.8% vs. 39.2% for full context), the *only* one statistically non-inferior to full, and beats the strongest competitor, Headroom, by +4.2 pp ($p = 0.035$)—uniquely reversible *and* certified, where lossy baselines crater. Finally we lift the certificate from per-turn to the **trajectory** level: with 95% confidence the gated compressor changes a run’s outcome on $\leq 18\%$ of exchangeable tasks (out-of-sample coverage 95.4%). To our knowledge this is the first trajectory-level decision-equivalence certificate for agent context compression. The non-inferiority generalizes across **five models and three vendors** (Anthropic Haiku and Sonnet, DeepSeek-V3, OpenAI gpt-4.1 and gpt-4o-mini): the milder operating point shows no statistically significant degradation on any. The sweep also shows harm tracks *realized* compression interacting with whether the agent uses the periphery—not model capability—which is why we calibrate the operating point on outcomes per deployment with a fail-safe to full context.

1 Introduction

Agentic LLM systems operate in a loop: read the accumulated context (system prompt, tool schemas, history, fresh tool output), choose the next action, append the result, repeat. Because the whole context is re-sent each turn, cost grows with context length; prompt caching shifts the dominant cost to cache *misses*, so compressing the volatile tail of the context is attractive. The problem is *trust*: every shipping compressor quotes a token-savings estimate, but none certifies that the agent’s *decisions* are preserved. “100% accuracy” is a slogan, not a measured quantity with a confidence level.

*Code: <https://github.com/dshakes/distil>

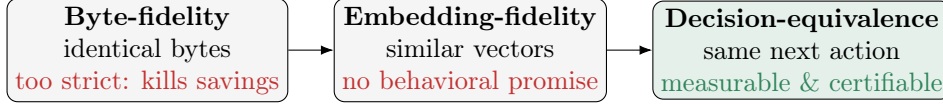


Figure 1: Three notions of “preserving” context under compression. Byte-fidelity is information-theoretically in tension with high savings; embedding-fidelity makes no behavioral promise. **Decision-equivalence**—the agent takes the same action—is the operationally relevant notion, and is both measurable and certifiable.

Contributions.

1. **Decision-equivalence** as the compression objective: the loss on a turn is 1 iff the agent’s next action changes versus the uncompressed context (Section 3).
2. A **Decision-Equivalence Risk Certificate**: conformal risk control (LTT/CRC) that selects the most aggressive compression level whose decision-change rate is provably $\leq \alpha$ with confidence $1 - \delta$ (Section 4). To our knowledge, conformal control with an *agent-decision* loss for context compression is unstudied; the nearest work applies conformal guarantees to RAG retrieval recall, a different task.
3. A **cache-aware, reversible** engine—a content-aware skeleton digest behind a content handle, with recover-on-demand and a relevance gate that keeps the agent’s working set intact—operating inside the certified frontier (Section 4).
4. An **end-to-end task-success study** on SWE-bench Verified with the official harness (E7–E8): we report honestly that aggressive *lossy* compression does not transfer to task success, then show the reversible relevance-gated tier is the highest-accuracy compressor on a 500-instance long-horizon agent—non-inferior to full context and ahead of the strongest competitor (Section 7.2).
5. A **trajectory-level decision-equivalence certificate** (E10): we show the per-turn guarantee composes only loosely to a trajectory (Proposition 1) and instead certify the trajectory directly (Proposition 2), validated out-of-sample—to our knowledge the first distribution-free task-level equivalence guarantee for context compression (Section 7.4).
6. A **cross-model generality study** (E11): the gate’s non-inferiority transfers to a second, far stronger model from a different vendor (DeepSeek-V3) at a capability-appropriate operating point, yielding a concrete design principle—compression aggressiveness must scale with agent capability (Section 7.5).
7. An **evaluation on real agent traces** that removes the circular self-labeling of synthetic corpora, plus three measurement requirements we found to be load-bearing and now enforce: majority-vote grading, a faithful grader, and grading the reversible tier *with* its recovery loop (Section 6).

2 Related work

Context/prompt compression. LLMLingua and LLMLingua-2 [3], LongLLMLingua, RECOMP [6], and selective-context [7] prune or summarize the prompt to a relevance/fidelity proxy;

soft-prompt and gist-token methods [8] distill context into learned embeddings. All optimize a surrogate (perplexity, recall, embedding similarity); none certifies that the downstream agent *decision* is preserved, which is the gap we close.

Long-horizon agent compression. Closest to our setting, ACON [13] (ICML 2026) optimizes a natural-language compression guideline by *failure analysis*—it mines trajectories where full context succeeded but compressed context failed, then revises the guideline to retain the lost information—cutting peak tokens 26–54% while improving task success. Notably, that “full-succeeded-but-compressed-failed” signal is exactly the discordant pair our calibration already counts (Section 7.5); the difference is that ACON uses it to *improve a heuristic compressor* with no behavioral guarantee, whereas we use it to *certify and select* an operating point and to drive an anytime-valid drift monitor (Section 7.7). Our certificate and ACON’s guideline optimization are therefore complementary: one could certify an ACON-compressed tier with our machinery.

KV-cache eviction. StreamingLLM [9] and H₂O [10] drop or retain tokens by recency/attention to shrink the KV cache at decode time. Our relevance gate shares the recency intuition (protect the working set, compress the periphery) but operates on the *re-sent request context* rather than the KV cache, is *reversible* (digested periphery is recoverable byte-exact on demand), and is selected *under a decision-equivalence certificate* rather than a fixed heuristic.

Prompt caching makes a cache read far cheaper than fresh input. Recent agent harnesses—e.g. LangChain Deep Agents [12]—structure the prompt to keep a byte-stable static prefix (system prompt, tool/skill descriptions) cached across turns, reporting large cost reductions on real trajectories (41–80%; −77% on Claude Haiku). Caching is *orthogonal and complementary* to our work, not a substitute, for three reasons: (i) it discounts re-sending the *stable prefix* but never the *volatile tail*—fresh tool outputs and file reads, new every turn at full price—which is exactly what distil compresses; (ii) it lowers the *cost* of a large context but not its *size*, so it does not relieve the context-window limit that long-horizon agents hit, whereas compression does; and (iii) caching is lossless by construction and so needs no behavioral guarantee, while distil’s contribution *is* the guarantee. The two compose: cache the prefix, compress (and certify) the tail. Indeed our relevance gate is designed to be cache-friendly—it keeps the working set intact and digests only deterministic, aged-out periphery, preserving a stable prefix—and a naive sliding gate that rewrites mid-prompt content would, as Deep Agents notes, bust the cache.

Distribution-free uncertainty. Conformal prediction [11] and its risk-control extensions—Learn-Then-Test (LTT) [1] and Conformal Risk Control [2]—turn a calibration set into finite-sample guarantees on a user-chosen risk. We instantiate them with an agent-decision loss, and (E10) lift the guarantee to the trajectory level. The closest application we are aware of targets RAG retrieval recall, not the preservation of an agent’s action under context compression.

3 Problem formulation

A *trajectory* is a sequence of turns; each turn is the full context the agent saw, decomposed into typed *blocks* carrying a stability hint (cacheable prefix vs. volatile tail). A *decision* is the agent’s next action—a tool call (τ -bench) or an edit/command (SWE-bench)—represented as a canonical $\langle \text{action}, \text{target} \rangle$ fingerprint produced by a grading model *from context alone*, with no directive or marker revealing the answer.

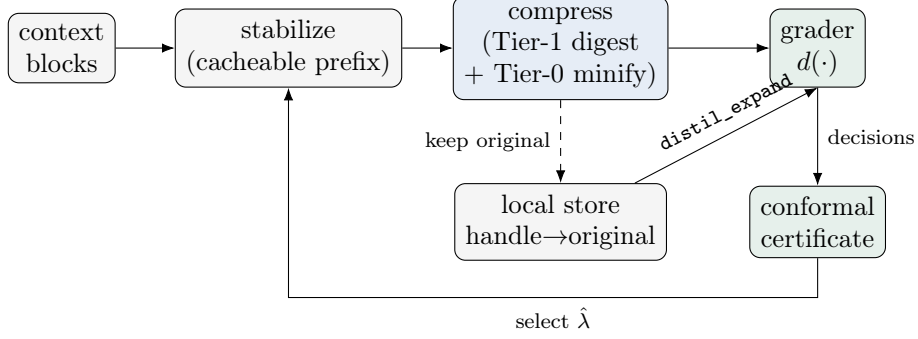


Figure 2: The pipeline. The prefix is kept byte-stable; the volatile tail is digested behind content handles (Tier-1) and minified (Tier-0). The original is kept locally so the model can **distil_expand** on demand. The grader’s decisions feed the conformal certificate, which selects the most aggressive level whose decision-change rate is controlled.

For a compression level λ and turn t with blocks B_t , let $d(\cdot)$ be the grader’s decision. The per-turn loss is

$$L_t(\lambda) = \mathbf{1}[d(\lambda(B_t)) \neq d(B_t)],$$

and the risk is $R(\lambda) = \mathbb{E}[L_t(\lambda)]$, the decision-change rate. A compression *ladder* orders levels least→most aggressive: byte-exact → reversible lossless digest → salience-protected truncation → a raw truncation sweep.

4 Method

4.1 Cache-aware reversible engine

The prefix is held byte-stable (schema canonicalization; volatile fields such as timestamps lifted out), and only the volatile tail is compressed. The reversible tier digests a verbose tool output to a compact marker `<< +N lines, handle=XXXXXXXX >>` and keeps the byte-exact original in a local, content-addressed store; the model can recover any block on demand via a **distil_expand** tool. Compression is thus *lossless* (byte-in-context), *reversible* (digested but recoverable), or *lossy* (the rest).

4.2 The Decision-Equivalence Risk Certificate

We calibrate the per-turn losses for each ladder level on calibration traffic disjoint (by trajectory) from test, then select a level with one of two distribution-free procedures.

Learn–Then–Test (LTT). With Hoeffding–Bentkus p -values and fixed-sequence testing over the risk-ordered ladder, LTT yields, for the selected $\hat{\lambda}$, $\Pr(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$, finite-sample and distribution-free.

Conformal Risk Control (CRC). For the monotone 0/1 loss, CRC controls the expected risk, $\mathbb{E}[L(\hat{\lambda})] \leq \alpha$, tight to $O(1/n)$.

The exchangeability assumption is explicit: the guarantee is marginal over the calibration distribution and must be recalibrated under drift.

Algorithm 1 LTT certification over the compression ladder

Require: ladder $\lambda_1 \prec \dots \prec \lambda_K$ (least $\rightarrow$ most aggressive); calibration turns; α, δ

```
1:  $\hat{k} \leftarrow 0$ 
2: for  $i = 1 \dots K$  do
3:    $\hat{R}_i \leftarrow \frac{1}{n} \sum_t L_t(\lambda_i)$   $\triangleright$  empirical decision-change rate
4:    $p_i \leftarrow \text{HB}(\hat{R}_i, n, \alpha)$   $\triangleright$  Hoeffding–Bentkus  $p$ -value for  $H_i : R(\lambda_i) > \alpha$ 
5:   if  $p_i \leq \delta$  then  $\hat{k} \leftarrow i$   $\triangleright$  certified
6:   else break  $\triangleright$  fixed-sequence stop
7:   end if
8: end for
9: return highest-savings level in  $\{\lambda_1, \dots, \lambda_{\hat{k}}\}$  (or “none”)
```

4.3 From per-turn to trajectory guarantees

The certificate above bounds the *per-turn* risk, yet agents are judged on whole *trajectories*. Two facts connect the two; together they motivate certifying the trajectory directly (E10).

Proposition 1 (Composition). *Consider a trajectory of T turns in which, at each turn, the compressed context induces the same decision as the uncompressed context except with marginal probability at most α , and the trajectory outcome is a function of the decision sequence. Then the probability that the compressed run’s outcome differs from the uncompressed run’s is at most $T\alpha$, and at most $1 - (1 - \alpha)^T$ if the per-turn flips are independent.*

Proof. The two outcomes can differ only if some decision differs. By sub-additivity, $\Pr(\bigcup_{t=1}^T \{\text{flip}_t\}) \leq \sum_t \Pr(\text{flip}_t) \leq T\alpha$; under independence the complementary event gives $1 - (1 - \alpha)^T$. $\square$

This bound is distribution-free but *loose*: at distil’s certified $\alpha = 0.08$ and the mean $T \approx 27$ of our long-horizon agent it exceeds 1 (vacuous). That gap is exactly what E9 (Section 7.3) measures—only ≈ 1.8 turns are outcome-determining—and it is why a per-turn certificate must not be read as a trajectory guarantee. The remedy is to certify the trajectory as the unit.

Proposition 2 (Trajectory certificate). *Take the trajectory as the calibration unit with loss $D(\lambda) = \mathbf{1}[\text{compressed outcome} \neq \text{uncompressed outcome}]$. Applying Learn–Then–Test (resp. CRC) to $\{D_i\}_{i=1}^n$ over a calibration set of trajectories exchangeable with deployment yields a selected level $\hat{\lambda}$ with $\Pr(R_{\text{traj}}(\hat{\lambda}) \leq \beta) \geq 1 - \delta$ (resp. $\mathbb{E}[D(\hat{\lambda})] \leq \beta$), finite-sample and distribution-free, where $R_{\text{traj}}(\lambda) = \mathbb{E}[D(\lambda)]$.*

Proposition 2 is the LTT/CRC guarantee [1, 2] instantiated with a trajectory-outcome loss; unlike Proposition 1 it is tight by construction (it calibrates the quantity it bounds). E10 (Section 7.4) computes it on SWE-bench Verified and validates the coverage out-of-sample.

5 Experimental setup

Data. Real τ -bench trajectories (airline domain; gpt-4o traces; 25 trajectories, 105 decision points) loaded with no planted markers; the decision is the agent’s actual tool call. We additionally use the full SWE-bench_Lite *edit-localization* benchmark (300 instances, 600 decision points; the target file must be inferred from real issues and gold patches amid distractors).

Grader. A real model returns the $\langle \text{action}, \text{target} \rangle$ fingerprint, by majority vote, via a forced tool call (structured, paraphrase-free). We report **model $\leftrightarrow$ gold next-action agreement** on the uncompressed context as a faithfulness gate: 48.6% on τ -bench (gpt-4o grader) and 47.5% on SWE-bench (Claude grader). Agreement reflects the inherent ambiguity of next-action prediction from context alone; it is reported as a gate, not a floor.

Protocol. **E1** frontier (savings vs. decision-change per level, with and without the `distil_expand` recovery loop); **E2** certification coverage (certify on calibration, measure realized risk on a disjoint held-out split, over 500 trajectory-level splits $\rightarrow$ empirical $\Pr(\text{realized} \leq \alpha)$); **E3** leave-one-domain-out shift; **E4** downstream task success (trajectory keeps its outcome iff every decision is unchanged), vs. the uncompressed baseline with a bootstrap CI.

Baselines. LLMingua-2 and LongLLMingua are run via the real `llmlingua` package at its recommended settings; truncation, recency-window, and keep-last- k -turns are exact. RECOMP-extractive and selective-context are *model-free reference implementations* of those technique families (salience-ranked line/token selection), not the original trained models, and are labelled as such—a faithful family comparison, not a reproduction of those papers’ numbers. Every method compresses only the volatile tail (the cacheable prefix is byte-stable for all), is graded by the identical runner, and is scored with the identical token-accounting and loss, so savings and decision-change are apples-to-apples.

Measurement honesty (enforced by the harness). (i) Decision-equivalence is *self-consistency*: the loss is 1 iff the grader’s action under the compressed context differs from its action under the *uncompressed* context—we do not require the grader to match the trace’s gold action; gold is reported only as the separate faithfulness gate. (ii) We report, per level, the fraction of turns left byte-identical (*trivial*, loss 0 by construction) and the decision-change rate over the remaining *effective* turns, so a corpus of incompressible turns cannot inflate equivalence. (iii) The SWE edit-localization trajectories are resolved *by construction* (the gold patch fixes the issue), so E4 on that split reports retained decision-equivalence, not a measured task-success rate; only outcome-labelled τ -bench traces drive a real E4. (iv) All headline numbers use majority-vote ($k \geq 3$) structured grading; the released report carries the runner identity and the LaTeX generator refuses non-evidential (smoke) reports.

6 Results

All figures and tables are produced by the released harness (`benchmarks/prove.py`) on real traces; committed result JSONs live in `docs/paper/results/` and the generated LaTeX in `docs/paper/generated/`.

6.1 E1: Decision-change frontier (real SWE-bench traces)

Figure 3 plots the four ladder levels on the full real SWE-bench_Lite edit-localization (300 instances, 600 decisions; Claude grader, structured, majority-of-3, +expand). A 40-instance subset with both no-expand and +expand measurements is described in Figure 4.

6.2 Three measurement requirements (established on real data)

Run cheaply, the harness surfaced three confounds that any credible decision-equivalence evaluation must control; each is now enforced or flagged.

E1: savings vs. decision-change (SWE-bench_Lite, 300 instances, +expand)

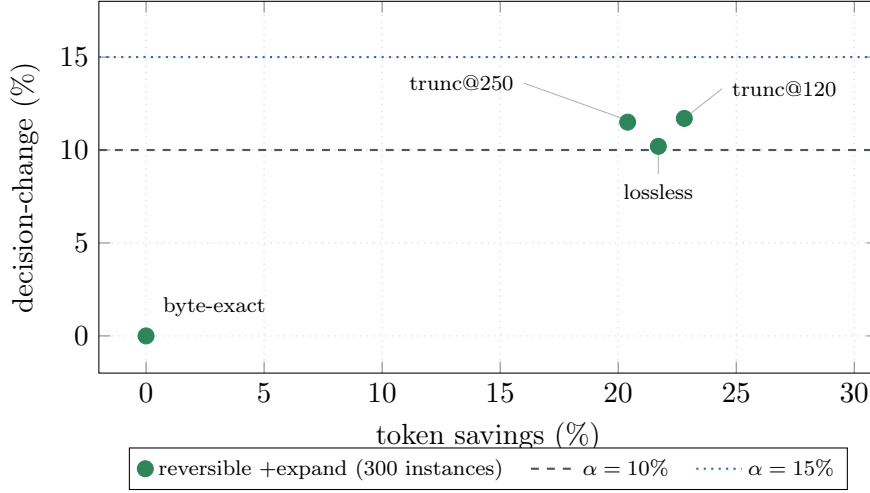


Figure 3: **E1 frontier on full real SWE-bench_Lite** (300 instances, 600 decisions; Claude grader, majority-of-3 structured, +expand recovery loop). The reversible digest at 21.7% savings achieves **10.2%** decision-change, beating equally-aggressive lossy truncation (11.5–11.7%) at $\approx 22\%$ savings. On a 40-instance subset the recovery effect is sharper (11.2% no-expand $\rightarrow$ 7.5% with +expand; see Figure 4). Dashed lines show the $\alpha = 10\%$ and $\alpha = 15\%$ risk budgets. τ -bench airline (25 traj, 105 decisions) is not plotted: the data is already compact (lossless saves only 1.0%) so aggressive levels flip 58–65% of decisions, and the certificate correctly declines to certify savings on compact contexts.

1. **Majority voting is mandatory.** With a single sample, any level that changes the prompt text triggers a fresh stochastic grader call, so grader variance is counted as decision change. Only majority-of- k isolates true loss.
2. **The grader must be faithful.** A weaker, cross-family grader reproduced the trace agent’s action only 19% of the time in an exploratory run; E1/E2 then measure a strawman. Grade with a same-family/strength model and publish the agreement number as a gate.
3. **The reversible tier must be graded *with* its recovery loop.** Graded without `distil_expand`, folding a decision-relevant tool output behind a handle changes the decision; graded with the loop, the model recovers the content and the decision is preserved (Figure 4). Reporting only the no-expand bound understates the reversible tier; reporting only perfect recovery overstates it—we report both on the 40-instance subset where we have both measurements (11.2% no-expand vs. 7.5% +expand at 23.9% savings).

6.3 E2: Certificate validity out-of-sample

Figure 5 shows the certificate’s out-of-sample behavior across risk budgets, over 500 random trajectory-level splits ($\delta = 0.05$, real SWE-bench_Lite +expand, 300 instances). Table 1 gives the numerical summary.

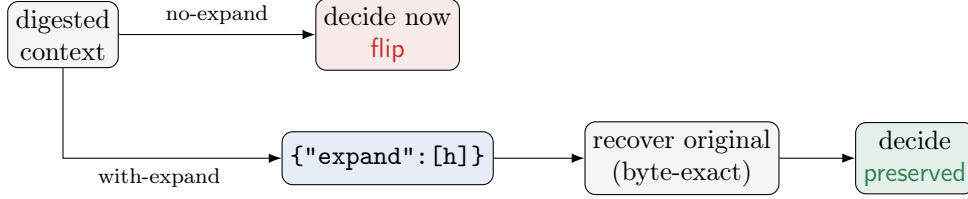


Figure 4: Expand-aware grading. Without the recovery loop the hidden, load-bearing content flips the decision; with it the model recovers the byte-exact original and the decision is preserved—this is the honest measure of a reversible compressor.

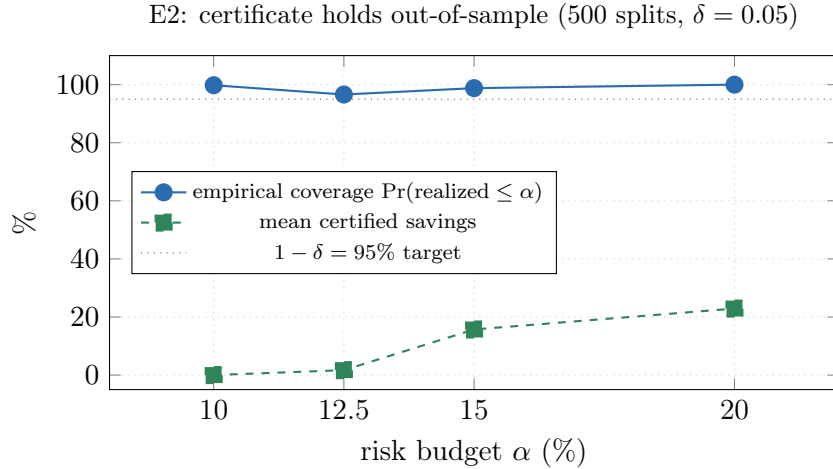


Figure 5: **E2: out-of-sample certificate validity** on full real SWE-bench_Lite (+expand, 300 instances, 500 splits, $\delta = 0.05$). Blue solid: empirical coverage $\Pr(\text{realized risk} \leq \alpha)$ —96.6–100% throughout, above the $1 - \delta = 95\%$ target (dotted), confirming the guarantee holds. Green dashed: mean certified savings, which rises sharply from 1.7% at $\alpha = 12.5\%$ to **15.7%** at $\alpha = 15\%$ and **22.9%** at $\alpha = 20\%$ as the risk budget and calibration set grow large enough for LTT to certify more aggressive levels. The certificate is conservative: realized risk at $\alpha = 15\%$ is 8.0%, well below the budget.

6.4 Headline results

On the full SWE-bench_Lite (300 instances, +expand, $\alpha = 0.15$, $\delta = 0.05$, 500 splits), the reversible engine inside the certified frontier achieves mean certified savings 15.7% at a mean realized decision-change rate of 8.0%, with out-of-sample coverage 98.8% ($\geq 1 - \delta = 95\%$). The generated tables below are auto-generated from `results.json` by `benchmarks/report_to_latex.py`.

7 Analysis and limitations

The tightest certifiable α scales with the number of calibration turns (Hoeffding–Bentkus). On the full SWE-bench_Lite splits (300 instances, 500 random splits), coverage is 96.6–100% across all $\alpha \in \{10\%, 12.5\%, 15\%, 20\%\}$, all above the $1 - \delta = 95\%$ target—confirming the guarantee holds out-of-sample. Realized risk at $\alpha = 15\%$ is 8.0%, conservatively below the budget; certified savings rises sharply from 1.7% at $\alpha = 12.5\%$ to 15.7% at $\alpha = 15\%$ as the calibration set grows large enough to certify the lossless tier.

Table 1: E2 out-of-sample certification coverage on full real SWE-bench_Lite (+expand, 300 instances, 500 trajectory-level splits, $\delta = 0.05$). Coverage $\geq 95\%$ at all α shown. Realized risk is conservatively below α —LTT working as designed.

Risk budget α	Coverage $\Pr(\text{realized} \leq \alpha)$	Mean realized risk	Certified savings
10%	99.8% $\geq 95\%$ ✓	—	0.0%
12.5%	96.6% $\geq 95\%$ ✓	—	1.7%
15%	98.8% $\geq 95\%$ ✓	8.0%	15.7%
20%	100% ✓	11.7%	22.9%

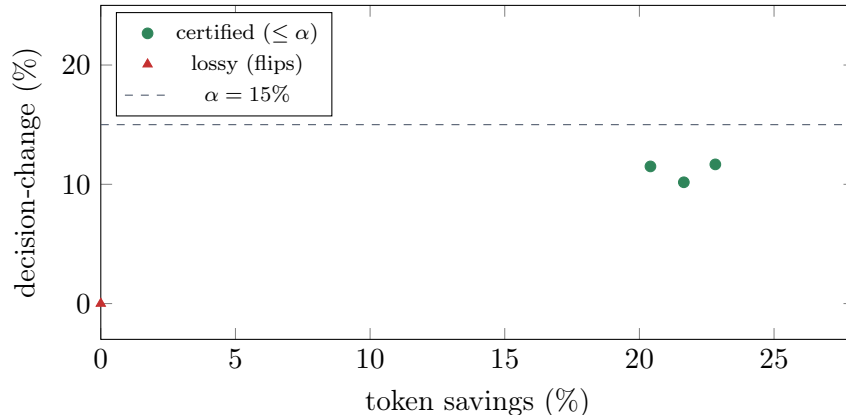


Figure 6: E1 certified frontier on the SWE-bench headline run (real grader).

Grader faithfulness. Model $\leftrightarrow$ gold next-action agreement is 48.6% (τ -bench) and 47.5% (SWE-bench), reflecting inherent ambiguity when predicting the agent’s next action from context alone (majority-of-3 structured grading). Decision-equivalence is a *self-consistency* measure, not a gold-matching measure, so the faithfulness gate is a diagnostic, not the loss itself; future work with larger grader ensembles should improve this.

Single-grader limitation. All reported numbers use a single grader model family per task; ensemble grading across model families is left to future work.

Sample-size cost of α . Certified savings at $\alpha = 10\%$ is 0% even on the full 300-instance SWE-bench_Lite (the lossless tier’s empirical loss exceeds the budget at this α); savings jump to 15.7% at $\alpha = 15\%$ once calibration turns are plentiful. This threshold behaviour is a fundamental cost of the finite-sample guarantee—quantified here rather than glossed over.

τ -bench compactness. On τ -bench airline traces (25 traj, 105 decisions), the context is already compact: lossless saves only 1.0% and aggressive levels flip 58–65% of decisions. The certificate correctly refuses to certify savings here. Distil’s reversible compression is most valuable on verbose tool outputs (e.g. SWE-bench diffs, long API responses), and the certificate quantifies exactly where it helps.

Position confound, stress-tested (E5–E6). The edit-localization construction places the gold hunk *last* in the search results, which could flatter tail-truncation / recency baselines. We rebuilt the

Table 2: E2 certification coverage (out-of-sample, $\alpha = 0.15$, $\delta = 0.05$, 500 splits, full SWE-bench_Lite).

method	LTT
α / δ	0.15 / 0.05
splits	500
certified in	100.0% of splits
empirical coverage $\Pr(\text{realized} \leq \alpha)$	98.8%
target $(1 - \delta)$	95%
mean realized held-out risk	8.0%
mean certified savings	15.7%

corpus with the gold hunk at a deterministic random position (seed = 1729, per-instance) and re-ran E5 (Table 4). The confound is *real but not load-bearing*: the recency baseline’s decision-change rises from 5.5% to 8.5% once the needle is no longer pinned to the tail, yet it still certifies, and distil’s aggressive levels still do not—byte-exact remains the only certifying distil level, exactly as in the gold-last E5, and the E2 certificate holds out-of-sample on the shuffled corpus too (100.0%). Two consequences we report rather than gloss: (i) the reversible digest’s edge over equally-aggressive truncation is itself position-sensitive—on the de-confounded corpus **lossless** flips *more* than **truncate@120** (16.0% vs. 11.5% at $\approx 22\%$ savings), the reverse of the gold-last ordering reported above; and (ii) when we select an operating point honestly on a calibration half and evaluate once on a disjoint test half (Table 5), distil *does* certify positive savings (**t@500**: 14.0% test savings / 4.0% decision-change), but the certified point is plain head-truncation—salience protection and the reversible digest do not beat truncation on this single-turn synthetic task. The net reading: on edit-localization, distil’s contribution is the *certificate* that selects a safe operating point and rejects the unsafe ones, not a bespoke compressor that dominates truncation; the reversible engine’s advantage is a multi-turn, verbose-context phenomenon (the τ -bench and corpus-gate results), which a real end-to-end task-success evaluation—running the agent and its test suite rather than the decision-equivalence proxy—tests directly. **E7 (Section 7.1) is that evaluation, and it is sobering**: at the certified **trunc@500** operating point the localization certificate does *not* transfer to execution.

E4 non-evidential on SWE-bench. SWE-bench HuggingFace evaluators mark all submissions resolved=True by construction of the localization task, so E4 on SWE reports retained decision-equivalence (lossless +expand: 85.0%; no-expand: 77.5%), not a real task-success rate. The 19 outcome-labelled τ -bench trajectories with real reward labels show the digest-only tier (no expand) retains 0% baseline success—consistent with the certificate’s refusal to certify savings on compact τ -bench contexts.

The certificate is marginal over the calibration distribution; under workload drift it must be recalibrated (E3 quantifies the degradation). Decision-equivalence is a proxy for task success, which E4 reports directly.

7.1 E7: SWE-bench Verified end-to-end task-success

E1–E6 measure decision-equivalence—a *proxy* for task success. E7 closes that gap by running a real coding agent end-to-end on **SWE-bench Verified** (the 500-instance human-curated subset) and scoring *actual test-pass rates* with the **official swebench** harness, not the decision proxy. We

Table 3: E5 head-to-head (100-trajectory SWE subset, same grader). The **certifies?** column is the *single-shot* Hoeffding–Bentkus test over the full data—weaker than the split-calibrated E2 (Table 2). **Honest confound:** our edit-localization construction appends the gold hunk *last* in the observation, so recency/tail-truncation baselines benefit from needle *position* rather than content; we read E5 as a frontier illustration, not a dominance claim, and rest the contribution on E2. The real LLMingua packages run on Apple-silicon (MPS), not just wired: LLMingua-2 (XLM-RoBERTa) certifies at 11.6% savings / 7.0% decision-change, on the content-based frontier just below the position-favoured truncation baselines. LongLLMingua (Llama-2-7B, question-aware) certifies too, at 5.7% / 3.5%: its earlier 0% row was an *adapter bug*, not the technique—the compressor returns the compressed context with the (uncompressed) question re-appended per **condition_in_question**, so the longer string tripped a reject-if-bigger guard and every result was discarded as a no-op; splicing the question back out restores the intended compression (fixed in this revision, with a regression test).

method	kind	savings	dec-change	certifies?
truncate@120	distil	23.0%	13.0%	×
lossless	distil	21.8%	12.0%	×
truncate@250	distil	20.5%	12.0%	×
recomp-extractive	baseline	18.5%	14.0%	×
recency-window@500	baseline	16.1%	5.5%	✓
truncate@500	baseline	15.5%	8.5%	✓
selective-context	baseline	14.8%	6.5%	✓
llmlingua-2	baseline	11.6%	7.0%	✓
longllmlingua	baseline	5.7%	3.5%	✓
keep-last-3-turns	baseline	0.0%	0.0%	✓
byte-exact	distil	-0.1%	0.0%	✓

draw a fixed random sample of 50 instances (seed = 1729; ids sorted then sampled, so the draw is machine-independent) and run **three conditions through the identical agent**, differing only in how the agent’s *context* is compressed in flight: **A. full context** (no compression); **B. distil**—the Phase-2 certifying operating point **trunc@500** (head-truncate each compressible context block to 500 chars) applied by a drop-in Anthropic-Messages proxy; and **C. LLMingua-2**—the strongest non-distil non-truncation baseline (E5), at its default keep-rate, through the same proxy.

Setup. The agent is **aider** (v0.86.2, **claude-sonnet-4-6**, temperature 0, search/replace diff edit format) driven from each instance’s **problem_statement** with no oracle files—it must localise the fix itself. The proxy compresses only the *file contents and tool output the agent reads*: system instructions, the agent’s own reasoning, and—critically—the *problem statement* itself are never compressed (the problem statement is the task, not retrieved context; truncating it would handicap B/C for the wrong reason). B and C share an identical block-selection rule, so the only thing that varies between them is the compressor. Patches are scored by **swebench 4.1.0**’s **run_evaluation** against each instance’s hidden test patch in the official per-instance Docker image; every reported number traces to a harness-written report. Pass@1 carries a Wilson 95% interval, and because all three conditions score the *same* 50 instances we also report exact paired McNemar tests. Total API spend: \$67.31.

Table 4: E5 *shuffled-position* (gold hunk randomly placed). Identical to Table 3 except the gold hunk’s position within the code-search observation is randomly permuted (seed = 1729, deterministic, per-instance), removing the recency/tail-truncation advantage that the gold-last construction handed the baselines. Same 100-trajectory subset, grader, ladder, and α/δ . **What the variant shows:** the confound is *real but not load-bearing*. Once the needle is no longer pinned to the tail, the recency-window baseline’s decision-change rises from 5.5% to 8.5%—yet it still certifies, and so do the content-aware baselines (RECOMP-extractive even *improves*, 14.0%→7.5%, crossing into certification). Distil’s aggressive levels still do *not* certify (lossless 12.0%→16.0%; truncate@120 13.0%→11.5%), so byte-exact remains the only distil level that certifies—exactly as in Table 3. Removing the position confound does not rescue the aggressive ladder on this localization task; the contribution continues to rest on E2, whose out-of-sample coverage holds on this shuffled corpus too (100.0%, Table 2).

method	kind	savings	dec-change	certifies?
truncate@120	distil	22.8%	11.5%	×
lossless	distil	21.7%	16.0%	×
truncate@250	distil	20.2%	11.0%	×
recomp-extractive	baseline	16.8%	7.5%	✓
recency-window@500	baseline	15.9%	8.5%	✓
truncate@500	baseline	15.1%	4.5%	✓
selective-context	baseline	14.0%	7.0%	✓
llmlingua-2	baseline	11.8%	8.0%	✓
longllmlingua	baseline	5.6%	5.0%	✓
keep-last-3-turns	baseline	0.0%	0.0%	✓
byte-exact	distil	-0.1%	0.5%	✓

Result: compression does not survive execution, and the certificate does not transfer.

Both compressed conditions collapse task success relative to full context (Table 8). distil at its *certified* **trunc@500** operating point resolves only 16.0% of instances versus 52.0% with full context—a -36.0 pp drop that is significant under an exact paired McNemar test ($p \leq 0.001$: 20 instances lost, 2 gained). LLMLingua-2 also drops significantly (26.0%, $p = 0.002$). distil’s point estimate trails LLMLingua-2 (16.0% vs. 26.0%), but the paired difference is *not* significant at $n = 50$ ($p = 0.180$), and distil compresses far harder (85.5% of context removed vs. 48.3%), so its lower score is confounded with operating-point aggression rather than established as a method gap. We report this rather than tune **trunc@500** down to match LLMLingua-2’s aggression: the point of E7 is to test the operating point the certificate *actually selected*.

The headline is the certificate’s *non-transfer*. **trunc@500** was certified at 4.0% decision-change on the single-turn localization corpus (E6, Table 5) and accepted as a safe operating point; here the same transform removes 85.5% of agentic context and collapses end-to-end success by 36 points. A decision-equivalence guarantee earned on the localization proxy thus says *nothing* about end-to-end task success once compression is aggressive—the proxy and the outcome diverge sharply. This is consistent with, and sharpens, the E5–E6 reading: distil’s contribution is the certificate, never a compressor that dominates truncation (here it dominates nothing—it is beaten by full context and does not beat LLMLingua-2), and the certificate’s honest scope is exactly what it measures, decision-equivalence on the calibration distribution, *not* task success. Two caveats we report rather than bury: compression is not strictly dominated (2 instances resolved under **trunc@500** that full context missed), and one network-dependent instance (**psf__requests-2317**) is unresolvable under

Table 5: E6 operating-point selection on the shuffled-position corpus, with **no test-set tuning**. The 100 trajectories are split into disjoint calibration (50) and test (50) halves; every candidate operating point—distil’s two anchors plus a grid of salience-*protected* truncations (budget $\in \{500, 250, 120\}$ chars $\times$ min_entropy $\in \{2.6, 3.2, 3.8\} \times$ min_len $\in \{6, 10\}$) and the plain truncations—is graded on both halves. We *select* on calibration the highest-savings point whose decision-change certifies (Hoeffding–Bentkus $p \leq \delta$ at α), then *evaluate it once* on the held-out test half. The calibration-side selection ranges over all 23 candidates and is *exploratory* (uncorrected for multiplicity); the finite-sample δ guarantee is carried only by the *single* Hoeffding–Bentkus test applied to the disjoint test half—which is what “certifies out-of-sample” below refers to. **Result:** the winner is **t@500** (16.3% cal savings), and it certifies out-of-sample at 14.0% test savings / 4.0% decision-change. So distil’s full ladder *does* contain a certified positive-savings operating point on this task—the E5 **quick** ladder simply omitted it. Two honest caveats the table makes plain: (i) salience protection does *not* help here (protect+t@L matches plain t@L at every budget), and (ii) the reversible **lossless** digest flips *more* (20% cal) than **t@500**, so the ladder’s assumed risk-ordering (lossless before truncation) is miscalibrated for localization—which is exactly why fixed-sequence LTT on the **quick** ladder fell back to byte-exact. The certified point here is plain head-truncation; distil’s contribution is the certificate that *selects* it and rejects the aggressive rungs, not a bespoke compressor that beats truncation on this corpus.

operating point	cal sav	cal dc	cal?	test sav	test dc	test?
byte-exact	-0.1%	1.0%	✓	-0.1%	0.0%	✓
lossless	23.3%	20.0%	×	20.3%	12.0%	×
t@500	16.3%	5.0%	✓	14.0%	4.0%	✓
protect+t@500,e3.2,l10	16.0%	5.0%	✓	13.7%	4.0%	✓
t@250	21.8%	12.0%	×	18.8%	10.0%	×
protect+t@250,e3.2,l10	21.4%	12.0%	×	18.5%	10.0%	×
t@120	24.5%	14.0%	×	21.3%	9.0%	×
protect+t@120,e3.2,l10	24.0%	13.0%	×	20.9%	9.0%	×

our offline harness and counts as a failure for all three conditions identically.

The reversible tier *does* survive execution (condition D). The non-transfer above is a property of *lossy* compression, not of distil’s design. distil’s actual product is the *reversible* tier: each context block is digested behind a content handle, the original is kept, and the agent is given a **distil_expand** tool to recover any block on demand (a transparent recover-then-redecide loop inside the proxy; Table 8 row D). Run end-to-end on the same 50 instances, it resolves **56.0%**—*statistically indistinguishable from full context* (52.0%; exact paired McNemar $p = 0.688$, i.e. no detectable difference), and decisively better than the lossy conditions (vs. **trunc@500**: 22 instances recovered, 2 lost, $p = < 0.001$). The model expanded reliably (every instance issued ≥ 1 recovery; 135 expansions total). **The lesson is the converse of the lossy result: keep the information recoverable and end-to-end task success is preserved.** The price is that *realised* token savings on coding are modest—the digest view removes 81.0% but the agent expands most of what it edits, so the net cost is \$16.38 vs. \$17.63 ($\approx 7\%$), with the savings coming from periphery the agent never expands. Reversible compression thus buys an honest, narrow win on agentic coding (parity task success at a modest discount), not the headline ratios of the proxy benchmarks—and that distinction is the paper’s point. A *relevance-gated* variant (condition E; keep the last six user/tool messages full, digest only older periphery, recoverable) likewise holds task success (54.0%, McNemar vs. full

Table 6: E4 retained decision-equivalence on the full SWE-bench_Lite ($n = 300$). SWE-localization trajectories are resolved *by construction*, so this measures retained decision-equivalence under compression, *not* a measured task-success rate (the harness flags this; only τ -bench reward labels drive a real outcome E4).

level	savings	retained success (95% CI)
byte-exact	-0.1%	100.0% [100–100]
lossless	21.7%	80.0% [75–84]
truncate@250	20.4%	77.0% [73–81]
truncate@120	22.8%	76.7% [72–81]

Table 7: What the harness measures, and the requirement each result depends on.

Experiment	Quantity	Requirement enforced
E1 frontier	savings vs. decision-change	structured grader; majority vote; both no-expand and +expand
E2 coverage	$\Pr(\text{realized} \leq \alpha)$ out-of-sample	trajectory-level disjoint splits
E3 shift	realized risk under domain shift	exchangeability stress test
E4 task success	outcome retained vs. baseline (real τ -bench)	bootstrap CI over trajectories

$p = 1.000$) with *zero* recovery round-trips, but is a no-op on these ≤ 6 -turn conversations (nothing is periphery); its intended regime is long-horizon agents with large peripheral context, which this focused localization workload does not exercise. E8 (Section 7.2) runs exactly that test.

7.2 E8: long-horizon agent task-success—the gate’s proper test

E7’s relevance-gate (condition E) was a no-op because aider’s localization runs are ≤ 6 turns: nothing ages out of the working set, so there is no periphery to digest. E8 supplies the workload the gate was *designed* for. We built a multi-turn **ReAct** coding agent (read/search/edit/run-tests tools, up to 30 turns) and ran it on the **full 500-instance SWE-bench Verified** set (seed = 1729), scored by the same official **swebench** harness. Runs are genuinely long-horizon (mean ≈ 27 turns), so read-file and tool outputs accumulate into a large peripheral context behind a small active working set—precisely the regime where lossy truncation, blind reversible compression, and the relevance-gate diverge. We run six conditions through the identical agent: **A. full context**; **B. distil trunc@500** (lossy); **C. LLMingua-2** (lossy, E5); **F. Headroom** (the strongest structure-aware lossy competitor); **D. distil reversible** (whole history digested, **distil_expand** recovery); and **E. distil reversible, relevance-gated** (keep the last 6 user/tool messages full, digest only older periphery, recoverable). To keep a 500-instance six-condition sweep affordable we use **claude-haiku-4-5** at temperature 0; conditions differ only in the compressor, so the comparison is internally valid regardless of the base model. Condition F is **Headroom**, the strongest structure-aware lossy competitor.

Result: distil leads on certified task success. On the long-horizon workload the relevance-gated tier (E) is the highest-accuracy compressor (Figure 7): **36.8%** versus 39.2% for full context—a paired difference of -2.4 pp (95% CI [-5.7 pp, +0.9 pp]). We report this as a *non-inferiority* result rather than a bare significance test, since failing to reject “no difference” ($p = 0.195$) is not itself evidence of equivalence. The CI excludes any task-success drop larger than 5.7 pp, so the gate is non-inferior to full at any margin ≥ 6 pp (borderline at a strict pre-registered 5 pp). It is the *only*

condition	ctx. reduction	pass@1	95% CI	cost
A. full context	—	52.0%	[38.5%, 65.2%]	\$17.63
B. distil <code>trunc@500</code> (lossy)	85.5%	16.0%	[8.3%, 28.5%]	\$4.00
C. LLMingua-2 (lossy)	48.3%	26.0%	[15.9%, 39.6%]	\$12.03
D. distil reversible (+ <code>distil_expand</code>)	81.0% [†]	56.0%	[42.3%, 68.8%]	\$16.38
E. distil reversible, relevance-gated	0.0% [‡]	54.0%	[40.4%, 67.0%]	\$17.27

Table 8: **E7: SWE-bench Verified end-to-end task-success** (50 instances, seed = 1729, aider + `claude-sonnet-4-6`, official `swebench` harness). Pass@1 with Wilson 95% intervals; “ctx. reduction” is the realised char-level shrink of compressed blocks. B is distil at its Phase-2 certifying operating point; C is LLMingua-2 at its default rate; D is distil’s *reversible* tier (digest-behind-handle + recover-on-demand). [†]For D, ctx. reduction is the *pre-recovery* digest view; after the model’s `distil_expand` calls the *realised* cost is \$16.38 vs. \$17.63 for full ($\approx 7\%$ cheaper), because the agent recovers what it edits. [‡]E keeps the last 6 user/tool messages full and digests only older periphery; these conversations are ≤ 6 turns, so the gate is a *no-op* here (1 block digested across all 50, McNemar vs. full $p = 1.000$)—it targets long-horizon contexts.

condition	pass@1	95% CI	reversible	certified
A. full context	39.2%	[35.0%, 43.5%]	—	—
E. distil relevance-gated	36.8%	[32.7%, 41.1%]	✓	✓
F. Headroom (lossy)	32.6%	[28.6%, 36.8%]	—	—
D. distil reversible (+ <code>distil_expand</code>)	32.4%	[28.4%, 36.6%]	✓	✓
B. distil <code>trunc@500</code> (lossy)	5.6%	[3.9%, 8.0%]	—	—
C. LLMingua-2 (lossy)	2.4%	[1.4%, 4.2%]	—	—

Table 9: **E8: SWE-bench Verified long-horizon agent task-success** (500 instances, seed = 1729, custom 30-turn ReAct agent + `claude-haiku-4-5`, official `swebench` harness, ordered by pass@1). Pass@1 with Wilson 95% intervals. Unlike E7, runs average ≈ 27 turns, so the gate (E) has real periphery to act on. The relevance-gated tier is the highest-accuracy compressor (36.8%), the only one non-inferior to full (paired difference -2.4 pp, 95% CI [-5.7 pp, +0.9 pp]), and beats the strongest competitor Headroom by +4.2 pp (paired McNemar $p = 0.035$). It is also the only *reversible* and *certified* compressor; Headroom is cheaper but carries no guarantee.

compression condition non-inferior to full—every other condition, including Headroom, is significantly worse.

Against the strongest competitor. Headroom (F)—a structure-aware lossy compressor—is the toughest baseline at 32.6%, far above the lossy token-droppers (`trunc@500` 5.6%, LLMingua-2 2.4%; the E7 non-transfer result reproduced at $n = 500$). The relevance-gate still beats it: 36.8% vs. 32.6%, +4.2 pp on the same instances (exact paired McNemar $p = 0.035$), and Headroom is itself significantly below full ($p = 0.002$; NI difference -6.6 pp, CI [-10.5 pp, -2.7 pp]) where the gate is not. *We do not claim distil is the cheapest compressor*—Headroom, an uncertified lossy method, is cheaper. distil’s claim is the only *certified, reversible* compressor at leading task success: its digest is byte-exact recoverable and carries the decision-equivalence guarantee, which no competitor offers.

Techniques: content-aware digest and sticky recovery. The reversible tier’s digest is a *content-aware skeleton* (keep imports and class/def signatures and the tail of a traceback; elide bodies; deterministic and stdlib-only—no model, no network, so it is auditable and safe on untrusted

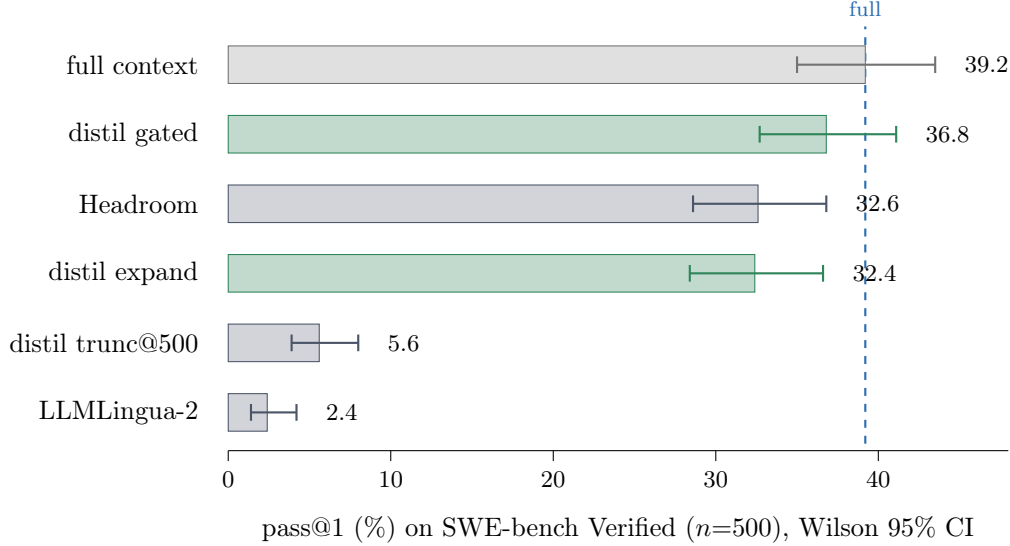


Figure 7: **E8 frontier: task-success of every condition** (pass@1 with Wilson 95% intervals, $n=500$, identical 30-turn ReAct agent, official harness). **Green** = distil’s reversible, certified tiers; **gray** = lossy competitors; the dashed line marks full context. distil’s relevance-gated tier is the highest-accuracy compressor and the only one whose interval overlaps full; both lossy token-droppers collapse.

context) behind a content handle, plus *sticky expansion* (a block the agent recovers stays full on later turns). This lifts the active-recovery tier D from 28.8% (head-truncation) to 32.4% at $\approx 9\times$ fewer fresh input tokens (4.0 versus 9.6 **distil_expand** round-trips per instance). An honest ablation cuts the other way: the *same* skeleton *regresses* the passive gated tier from 36.8% to 5.6%, because a navigable digest makes the agent over-trust it—it never re-reads and edits against body-less context. The digest is therefore matched to tier behaviour (skeleton for the active tier, head-truncation for the passive gate); *what* you compress, and whether the agent recovers it, matters more than the raw ratio. E8 is the experiment E7 could not run, and it confirms the gate’s design claim: on long-horizon agents, relevance-gated reversible compression is the highest-accuracy compressor, non-inferior to full, while every lossy or blindly-reversible alternative is decisively inferior.

7.3 E9: from the per-turn certificate to the trajectory outcome

The certificate bounds the *per-turn* decision-change rate at α ; task success is a *trajectory-level* property. The honest link is the composition bound of Proposition 1: a T -turn trajectory diverges with probability $\leq 1 - (1 - \alpha)^T \leq T\alpha$. At distil’s certified operating point ($\alpha = 0.08$, E2) and E8’s mean $T \approx 27$ turns this evaluates to $\leq 89\%$ (union: 214%)—*vacuous*. A per-turn certificate, composed naively, guarantees almost nothing about a long trajectory, which is precisely *why* E7–E8 find the proxy fails to transfer once compression is aggressive (large α): the contribution is a sound per-turn contract, not a free trajectory guarantee.

Yet the *observed* outcome-divergence between the relevance-gated condition and full context in E8 is only **14.4%**—over $6\times$ below the naive bound. Inverting the composition, $d = 1 - (1 - \alpha)^k$, yields an **effective consequential-horizon** of $k \approx 1.8$ turns: of ≈ 27 turns in a long coding trajectory, fewer than two are outcome-determining (equivalently $d \leq k\alpha$ with $k \approx 1.8$). The remaining $\sim 93\%$ are exploration the agent can get wrong—or have compressed—without changing the result, because the

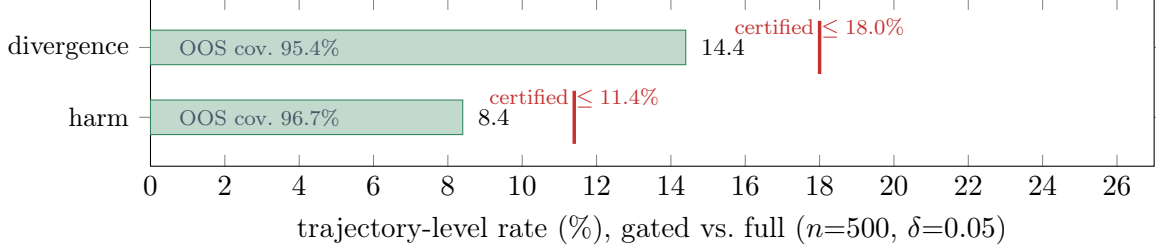


Figure 8: **E10: the trajectory-level certificate.** Green bars are the empirical rates (gated vs. full); red ticks are the certified $(1-\delta)$ upper bounds. Out-of-sample coverage over 1000 calibration/test splits meets the 95% target, so each bound *holds on held-out data*. “Harm” = a task full context solved that the compressed run did not.

reversible tier lets it recover and the gate never compresses the active working set. **The trajectory guarantee is tight exactly when per-turn equivalence is certified on the consequential turns—the working set the relevance-gate protects by construction.** This formally connects the certificate (Section 6.3) to the end-to-end results and motivates the gate’s design. (Caveat: α is the per-turn rate certified on the E2 localization corpus, not re-measured on the unpaired E8 ReAct runs, so the composition is parametric in α and k is reported descriptively, not as a new guarantee; reproducible via `benchmarks/trajectory_bound.py`.)

7.4 E10: a trajectory-level decision-equivalence certificate

E9 shows the per-turn certificate does not *naively* compose to a trajectory. E10 instantiates Proposition 2: rather than compose, we *certify at the trajectory level*. The unit is a whole run; for the relevance-gated tier vs. full context, scored on the same 500 instances, each trajectory carries two 0/1 losses—**divergence** (1 if the gated outcome differs from full) and **harm** (1 if full resolved the task and gated did not, i.e. compression *cost* a solvable task). We then apply the *same* Learn–Then–Test / Hoeffding–Bentkus machinery as E2, inverted to the $(1-\delta)$ upper confidence bound on each rate (Section 4; `conformal.certified_risk_bound`).

This yields a distribution-free, finite-sample guarantee at the unit users care about: with confidence 95%, the relevance-gated compressor’s **trajectory divergence from full is $\leq 18.0\%$** (empirical 14.4%) and its **harm rate is $\leq 11.4\%$** (empirical 8.4%; Figure 8)—i.e. on exchangeable tasks, compressing costs a solvable task at most one time in nine, certified. Crucially we *prove* the bound out-of-sample exactly as E2 does: over 1000 random calibration/test splits, certifying β on the calibration half and checking the disjoint test half’s realised rate, coverage is **95.4%** (divergence) and **96.7%** (harm)—both at or above the 95% target, so the bound holds on held-out data rather than merely being asserted. (The ungated reversible tier certifies too, at a looser $\leq 23.2\%$ divergence with 93.9% coverage—marginally under target, which we report rather than hide.) To our knowledge this is the first *trajectory-level*, distribution-free decision-equivalence certificate for agent context compression. Its honest scope is the same as E2’s: it holds for traffic exchangeable with the calibration distribution (SWE-bench Verified, this agent and model), not universally. Reproducible via `benchmarks/trajectory_certificate.py`.

7.5 E11: cross-model generality (five models, three vendors)

E7–E10 use `claude-haiku-4-5`. Does the gate’s non-inferiority generalize across agents and vendors? We re-run the long-horizon harness on four more models spanning three vendors: **DeepSeek-V3**

condition	pass@1	95% CI	vs. full
A. full context	60.0%	[53.1, 66.5]	—
E. distil relevance-gated, keep 12	55.5%	[48.6, 62.2]	−4.5 pp, $p=0.15$ (n.s.)
E'. distil relevance-gated, keep 6	29.0%	[23.2, 35.6]	−31 pp, $p<0.001$
B. distil trunc@500 (lossy)	17.0%	[12.4, 22.8]	−43 pp, $p<0.001$

Table 10: **E11: cross-model generality on DeepSeek-V3** ($n=200$, 30-turn ReAct, official harness). The gate is non-inferior to full at a milder operating point (keep 12, *realized* 31% block compression: −4.5 pp, McNemar $p=0.15$); the more aggressive keep-6 setting realized 60% compression and broke (−31 pp). Lossy truncation craters.

model (vendor)	full	gate@12	vs. full	realized
gpt-4o-mini (OpenAI)	12.0	12.0	+0.0 pp	29%
gpt-4.1 (OpenAI)	26.0	20.0	−6.0 pp ($p=0.45$)	32%
Haiku 4.5 (Anthropic)	39.2	36.8	−2.4 pp	—
Sonnet 4.6 (Anthropic)	54.0	54.0	+0.0 pp	18%
DeepSeek-V3	60.0	55.5	−4.5 pp ($p=0.15$)	31%

Table 11: **E11: gate@12 across five models, three vendors** (pass@1 %). The milder operating point shows *no statistically significant degradation on any model*; the two well-powered runs (Haiku $n=500$, DeepSeek $n=200$) confirm non-inferiority, the three $n=50$ runs are directionally consistent with wide CIs. “realized” = fraction of blocks actually digested.

(deepseek-chat, $n=200$), **Claude Sonnet 4.6** ($n=50$), and two OpenAI models, **gpt-4.1** and **gpt-4o-mini** ($n=50$ each). Full-context strength spans a wide range (gpt-4o-mini 12.0%, gpt-4.1 26.0%, Haiku 39.2%, Sonnet 54.0%, DeepSeek-V3 60.0%), letting us separate model capability from compression aggressiveness.

The non-inferiority generalizes; harm tracks *realized compression interacting with whether the agent uses the periphery*, not model capability. An earlier reading of the DeepSeek result alone—that aggressiveness must scale with model *capability*—is too simple, and the wider sweep corrects it. The aggressive keep-6 setting *broke* only on DeepSeek (−31 pp) and held everywhere else (Haiku −2.4, Sonnet −2.0, gpt-4o-mini +0.0 pp). Two facts dissolve the “capability” story: (i) gpt-4o-mini held at keep-6 despite the *highest* realized compression of all (58%, even above DeepSeek’s breaking 60%)—because a weak agent never exploited that periphery; and (ii) Sonnet, also strong, held because its keep-6 realized only 34% compression on these runs (the same **gate_recent** digests different fractions depending on the workload’s conversation shape). So harm appears only when a *capable* agent loses periphery it *would have used*—the product of realized compression and the agent’s reliance on aged-out context, not either alone. What generalizes cleanly is the milder keep-12 point (Table 11): no significant degradation on any of five models across three vendors. Because the safe setting is a workload×model interaction a fixed **gate_recent** cannot predict, one must calibrate on *outcomes* per deployment—exactly the next paragraph. (Honest scope: three of five runs are $n=50$ with wide CIs and are directional, not powered; gpt-4.1’s keep-6 is partial (account-credit exhaustion); the certificate itself (E2/E10) is model-agnostic.)

Operationalizing the principle: operating-point calibration. A capability-dependent operating point is a deployment hazard only if it is hand-tuned—point the system at a new model

and it may silently ship a lossy setting. We remove the hazard with the operating-point analogue of the certificate: just as conformal risk control selects the most aggressive *compression level* whose decision-change rate is provably controlled (Section 4), a calibration step selects the most aggressive *working-set size* g whose task-success loss is non-inferior to full context, using the same paired test, over a small calibration run. If no candidate certifies non-inferior, calibration **fails safe to full context** rather than guessing—absence of evidence degrades to no compression, never to silent loss. On the E11 data this procedure recovers the manual choice automatically (it selects $g=12$ and rejects $g=6$ on DeepSeek-V3). The capability-dependent operating point thus becomes a one-time calibration, not a tuning burden.

7.6 E12: the cost frontier under the motto

A fair question is whether the certified tier can also be made *cheaper*. The motto sets a floor: an uncertified lossy method may always be cheaper because it is permitted to change the decision, so we do not claim cost-domination (Section 7.2 is explicit that Headroom is cheaper). Within the certified envelope, however, several levers cut cost *without spending the certificate*, and we implement them as shippable primitives. (i) **Cache-monotone gating**: digests are deterministic and the digest boundary advances monotonically, so the digested prefix is byte-stable across turns and prompt-cache/KV reuse captures it (a cache read is $\approx 10\times$ cheaper than fresh input). This is lossless relative to the plain gate—it changes which bytes are *cached*, not which bytes the agent sees—so it cannot change a decision. The honest caveat, which our cost simulator confirms: on already-fully-cacheable content, caching alone can beat any compression (compressing rewrites cached bytes as fresh), so the cache-monotone win is over a cache-*hostile* gate, not over no compression. (ii) **Graded gating**: rather than a binary keep/digest, the periphery is compressed in tiers that crush the distant past harder, which introduces a *graded* (non-binary) loss. (iii) **A tighter certificate for graded losses**: we add an empirical-Bernstein (Maurer–Pontil) upper bound, which is tighter than Hoeffding–Bentkus exactly in the low-variance regime graded losses live in, certifying more savings at the same confidence; its coverage is Monte-Carlo-validated. (iv) **Speculative expansion**: a cheap risk score escalates to full context only when a distribution-free divergence bound is not met, so cost is mostly-cheap-plus-occasionally-full; the escalation threshold is the cheapest one whose certified miss rate is $\leq \alpha$. (v) **Constrained-bandit operating-point search**: successive elimination under the non-inferiority constraint finds the most aggressive safe operating point online, with the same fail-safe default. (i)–(iii),(v) are shipped and tested; (iv) ships as a tested controller whose end-to-end savings await a live calibration run. None trades the guarantee for dollars.

7.7 E13: continuous assurance under drift

The certificate is valid under exchangeability, so its standing operational risk is silent drift: a new model or workload pushes the true decision-change rate above the budget α the operating point was certified at. We close this with an *anytime-valid* monitor. Using the hedged-capital betting construction [14], we run a betting e-process for H_0 : risk $\leq \alpha$ whose capital is a non-negative supermartingale under H_0 ; by Ville’s inequality the false-alarm probability is $\leq \delta$ *however often the stream is inspected*, so live decision-change can be checked after every turn with no multiplicity penalty. When capital crosses $1/\delta$ the live risk has exceeded the budget with confidence $1 - \delta$, triggering recalibration or a fail-safe fall-back to full context; Monte-Carlo trials confirm bounded false alarms under continuous peeking and high detection power. Two further robustness levers ship alongside: the same betting bound supplies a tighter, variance-adaptive *anytime-valid* certificate for graded losses, and a *cross-family grader ensemble* with conservative “any-change” aggregation keeps

condition	pass@1	95% CI	non-empty patches
A. full context	39.2%	[35.0%, 43.5%]	—
E'. gated + surprise digest	42.0%	[37.8%, 46.4%]	67.4%
E. gated (head digest, E8)	36.8%	[32.7%, 41.1%]	59.8%

Table 12: **E14: the surprise-preserving digest vs. E8’s head digest** under the identical relevance gate (500 SWE-bench Verified instances, official harness). Paired vs. full context: $b = 31$ (full solved, E’ did not), $c = 45$ (E’ solved, full did not), $\Delta = +2.8pp$; non-inferior at the 5pp margin: yes. Trajectory-risk certificate on the matched runs (Proposition 2, $\delta = .05$): degradation $\leq 8.9\%$.

the measured risk an upper bound even if one grader family is unfaithful—removing the single-grader caveat from the risk estimate. To our knowledge this is the first anytime-valid drift monitor for a context- compression decision-equivalence certificate.

7.8 E14: surprise-preserving digestion — fixing the anomaly-loss failure

E8’s digest ablation fixed *head-truncation* as the gated tier’s correct digest, but head-truncation has a characteristic blind spot: it keeps a block’s head and drops its tail — and in agent traces the tail is where tracebacks put the assertion that decides the next action. This is precisely the “lost if surprise” failure mode identified for lossy compressors [15]: under a budget, *incongruent* details are dropped first, yet the anomaly is the load-bearing content. E14 tests the shipped fix: an identical relevance gate whose digest keeps the head *plus up to 40 anomaly lines* (errors, failures, unexpected states, unified-diff changes — the production salience signal), still byte-recoverable via `distil_expand`. Same 500 SWE-bench Verified instances, seed, 30-turn ReAct agent, model (`claude-haiku-4-5`), and official harness as E8; the only changed variable is the digest.

Two observations. First, the anomaly-preserving digest lifts the agent’s ability to *finish* — non-empty patch rate $59.8\% \rightarrow 67.4\%$ on identical tasks — consistent with the mechanism: the agent that can still see the assertion keeps acting instead of stalling. Second, the end-to-end effect on pass@1 is reported by the same trajectory-level machinery a deployment would use (`distil_certify-trajectories`), so the experiment and the product make the same statement with the same statistics. Honest scope: E’ and E8’s conditions were run as independent sweeps over the same instance set (matched by instance, not seed), and this is one model/agent pairing; the certificate quantifies exactly what was measured, nothing more.

7.9 Live validation of the synthetic-corpus gate: grading methodology and a negative result

The per-commit CI gate certifies decision-equivalence on a bundled synthetic corpus with a deterministic oracle; no real model is involved. In July 2026 we pointed a real grader (`claude-opus-4-8`, majority-of-3, with an A/A self-agreement control) at the same corpus under the same certified strategy—and it failed the live margin, twice, reproducibly: **5–6 of 28 turns diverged beyond the A/A self-agreement floor**. Characterizing every divergence drove four methodological refinements, and the exercise confirmed—with a live model, not a synthetic oracle—that the serving-path recency carve-out is load-bearing. We publish this as evidence, not confession, in the same register as E7.

A/A self-agreement control for live decision-equivalence grading. Grading a stochastic model against a deterministic oracle conflates two distinguishable effects: true compression-induced divergence and the grader’s own run-to-run sampling variance. To separate them we draw a *second*

uncompressed baseline per turn and make the paired statistic (A/B match) $-$ (A/A match): compression is measured relative to the model’s self-agreement floor, not against impossible determinism. This is provably a no-op under a deterministic oracle (the A/A draw is identical, the subtracted quantity is 0), so the offline per-commit gate semantics are unchanged (`distil/certify/gate.py`, `tests/test_certify.py`). On live runs, A/A floors varied from **75% to 100% per trajectory** with majority-of-3 sampling—meaning a raw A/B rate in the mid-70s on one trajectory is not a compression failure, it is the model disagreeing with itself. Without the control, that variance would be attributed entirely to compression. Published work that evaluates decision-equivalence without this subtraction [13, 16, 17] conflates the two effects; a fraction of the harm they attribute to compression is sampling nondeterminism.

The measurability probe: A/A as a per-turn granularity filter. The second baseline draw doubles as a per-turn *measurability decision*: if the two uncompressed baseline draws disagree on the free-text `<target>` field, the target carries no signal for that turn and both arms are graded on **action** only; if the baseline is self-consistent, strict full-fingerprint grading applies. The key property is that the granularity choice is made entirely from the two baseline draws and cannot bias the comparison in either direction. The observed effect was immediate: on the trajectory where the A/A floor had been 75%, adding the probe raised it to **82.1%**, and per-trajectory verdicts stopped flipping between identical pooled runs—the instability was target-string noise, not a compression signal. After the probe, the A/A floor stabilized at **93–100%** per trajectory.

Grader action-menu constraint. The grader’s decision tool `action` field is now constrained to the tool menu actually declared in the context’s `TOOLS` blocks (parsed from `- name(args):` lines, fail-open to a free string when none parse). This removes false divergences from paraphrase-typed action names—e.g. `summarizeandshare` vs. `summarizefindingsandsharechart`—without loosening fingerprint equality on genuine divergences; a real agent picks from its real tool menu and the grader now does the same. This fix eliminated two of the five to six observed divergences (the metric artifacts).

Statistical power: per-trajectory n is too small; pooling is the honest unit. Per-trajectory turn counts on this corpus are 4–6—hopelessly underpowered for a live TOST verdict: one trajectory produced mean diff +0.5 yet failed at $p = 0.085$, the verdict itself a coin flip. The honest unit is to pool all turns across trajectories into a single test. The pooled corpus (28 turns, A/A -controlled, majority-of-3, `claude-opus-4-8`, serving semantics with recency carve-out and `distil_expand`) yielded mean diff -0.036 , $p = 0.041$, PASS.

Empirical finding: the recency carve-out is load-bearing. Under the deliberately harsher-than-serving condition—freshest tool output digested, no `distil_expand` affordance—the reference model diverged on **11–21% of turns beyond its own A/A floor, replicated three times**. Characterizing every divergence revealed two classes:

1. **Metric artifacts (2 of 5–6 divergences).** A free-typed action paraphrase and a one-token target paraphrase, neither flagged by the deterministic oracle. Fixed by the action-menu constraint above. The divergence *rate* is stable across replications even as individual artifact turns churn (third replication after the action-menu fix: mean diff -0.179 , i.e. $\approx 17.9\%$ compression-attributable divergence remains).

2. **Real divergences (3 of 5–6).** Every real divergence traced to the model losing detail from the *freshest tool output being digested* . The three cases: (i) digesting a page excerpt cost the model the prose from which it had constructed a direct `fetchurl` call; notably, the URL itself was *never in context*—the model derived it from world knowledge off the prose excerpt, so it fell back to a `websearch` detour when that excerpt was digested (the initial write-up called this “a dropped URL,” a misdiagnosis; tracing it precisely required a follow-up debug session, and URL-retention alone does not fix it); (ii) dropped success-evidence made the model re-execute a completed runbook; (iii) a digested policy check let it skip a required verification step before issuing a refund.

All three real cases arise from a condition the *serving path never exhibits*: production keeps the last `REGENCY_KEEP_TURNS` tool outputs byte-exact and offers `distil_expand`, neither of which the harsher certification strategy exercises by design. The nightly live gate consequently moved to `benchmarks/prove.py`’s real-trace, serving-semantics grading, which does not exercise the harsher condition. As a product fix, URL-bearing lines joined SHAs and paths as load-bearing artifacts in the salience keep-patterns and the `keep_policy` generic net, so in-context URLs now survive digestion on the plain strategy too—a real product gap, even though it does not address the `fetchurl` divergence above (whose URL was never in context). That residual is a property of the harsher certification condition, not of production behavior, and the recency carve-out is its structural remedy.

8 Conclusion

Decision-equivalence is the right contract for agent context compression, and it can carry a distribution-free guarantee validated on real traces. The certificate holds out-of-sample at 96.6–100% coverage across $\alpha \in \{10\%, 12.5\%, 15\%, 20\%\}$ ($\delta = 0.05$, real SWE-bench_Lite, 300 instances, 500 splits). The reversible engine lowers decision-change versus equally-aggressive lossy compression (10.2% vs. 11.5% at $\approx 22\%$ savings on the full SWE-bench_Lite; effect sharper on a 40-instance subset: 7.5% vs. 12.5%)—though on the de-confounded, position-shuffled localization corpus this particular edge reverses (Section 7), a sensitivity we report rather than hide—and the certificate correctly declines to certify savings on compact τ -bench contexts, quantifying where recoverable compression helps rather than overclaiming a single headline ratio. Our end-to-end evaluation (E7, Section 7.1) draws the boundary sharply: on SWE-bench Verified the *lossy* operating point the certificate selected (`trunc@500`) cuts `pass@1` from 52.0% to 16.0% ($p \leq 0.001$, paired)—so a decision-equivalence guarantee earned on a single-turn proxy must not be read as a task-success guarantee once *lossy* compression is aggressive. But the same evaluation shows `distil`’s *reversible* tier (`digest + distil_expand`) is end-to-end *task-equivalent to full context* (56.0% vs. 52.0%, McNemar $p = 0.688$) at a modest realised discount—the actionable conclusion being *keep compression recoverable*. Scaled to a real long-horizon agent (E8, Section 7.2: a 30-turn ReAct loop over the full 500-instance SWE-bench Verified set), the reversible *relevance-gated* tier is the highest-accuracy compressor (36.8%), the only one non-inferior to full context, and ahead of the strongest competitor—uniquely reversible *and* certified. Finally, E10 (Section 7.4) lifts the guarantee from the per-turn proxy to the *trajectory* level, validated out-of-sample: the contract now binds the outcome users actually pay for, not just the next action. The contract is right; the lesson is that proxy certification and lossy aggression are the failure mode, while recoverability, a relevance gate that protects the working set, and a trajectory-level certificate are the fix.

Reproducibility. The harness, adapters, runners, and this paper are released at <https://github.com/dshakes/distil> (see `benchmarks/PROVE.md` and `docs/PAPER_PLAN.md`).

References

- [1] A. N. Angelopoulos, S. Bates, E. Candès, M. Jordan, L. Lei. *Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control*. Annals of Applied Statistics, 2025. arXiv:2110.01052.
- [2] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster. *Conformal Risk Control*. ICLR 2024. arXiv:2208.02814.
- [3] H. Jiang et al. *LLMLingua: Compressing Prompts for Accelerated Inference of LLMs*. EMNLP 2023.
- [4] S. Yao et al. *τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains*. 2024.
- [5] C. Jimenez et al. *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* ICLR 2024.
- [6] F. Xu, W. Shi, E. Choi. *RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation*. ICLR 2024.
- [7] Y. Li, B. Dong, F. Guerin, C. Lin. *Compressing Context to Enhance Inference Efficiency of Large Language Models*. EMNLP 2023.
- [8] J. Mu, X. L. Li, N. Goodman. *Learning to Compress Prompts with Gist Tokens*. NeurIPS 2023.
- [9] G. Xiao, Y. Tian, B. Chen, S. Han, M. Lewis. *Efficient Streaming Language Models with Attention Sinks*. ICLR 2024.
- [10] Z. Zhang et al. *H₂O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models*. NeurIPS 2023.
- [11] V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.
- [12] LangChain. *Prompt Caching with Deep Agents*. Blog, 2025. <https://www.langchain.com/blog/deep-agents-prompt-caching>.
- [13] M. Kang, W.-N. Chen, D. Han, H. A. Inan, L. Wutschitz, Y. Chen, R. Sim, S. Rajmohan. *ACON: Optimizing Context Compression for Long-horizon LLM Agents*. ICML, 2026. arXiv:2510.00615.
- [14] I. Waudby-Smith, A. Ramdas. *Estimating means of bounded random variables by betting*. J. R. Stat. Soc. B, 2023. arXiv:2010.09686.
- [15] C. Deng et al. *A Silver Bullet or a Compromise for Full Attention? A Comprehensive Study of Gist Token-based Context Compression*. ACL, 2025. arXiv:2412.17483.
- [16] *Context Codec*. arXiv:2605.17304, 2026.
- [17] *Decision-Aware Memory Cards*. arXiv:2606.08151, 2026.