ACTION-EQUIVALENT MEMORY SKETCHES FOR WEB AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Web agents powered by large language models often fail not because the underlying model is too weak, but because the memory that conditions the policy is semantically faithful yet action-irrelevant. We formalize this gap as an *action-equivalent memory sketch*: a compact history representation that preserves exactly the information capable of changing the next action, operationalized as an eight-field structured sketch (*goal*, *sub goal*, *completed*, *failed*, *constraints*, *entity bindings*, *pending*, *evidence*) updated at every step by a single LLM call. Three results bound its content: a soundness theorem for policy equivalence, an information-theoretic lower bound on sketch size, and a minimality theorem showing the eight fields are an action-complete cover.

We evaluate on Mind2Web *test website* across three runs. On *qwen-turbo* (10 task, 12 conditions) ours reaches Step-SR 0.0964; the B3 diagnostic, which removes *constraints* and *failed*, drops Step-SR to 0.0723, a 25% relative decrease consistent with the minimality theorem. On *GLM-5.3-Flash* (10 task, 8 per-field ablations) the direction reverses: B3 reaches 0.1233 versus ours 0.1096, with *sub goal* the largest up-move and *constraints* the only down-move, opposite to what the theorem predicts. Suspecting a sample-size artifact, we ran a 50-task replication restricted to B3 and ours (438 step-judgments per condition, horizons 2 to 16). The direction returns to ours above B3 at 0.1347 versus 0.1187, a 13.5% step-level gain. The two-backbone picture is that the sketch works on both backbones once the sample is large enough, and 10-task direction flips are sample-size artifacts.

1 INTRODUCTION

Web agents powered by large language models (LLMs) have moved from research curiosity to deployed products. The same failure recurs across systems: a model that successfully completes the first five steps of a twelve-step task suddenly repeats a step it has already taken, ignores a constraint stated in the original request, or "forgets" a failed attempt and re-executes it. Surveying public traces of Mind2Web (Deng et al., 2023), WebArena (Zhou et al., 2024), and SeeAct (Zheng et al., 2024) runs, we observe that these failures share a common cause: the memory representation that conditions the policy is semantically plausible but not *action-equivalent*.

Three memory mechanisms dominate current web-agent implementations. The first is full history: concatenate every prior observation, action, and result. This is the natural upper bound, but it is expensive, slow, and saturates the context window long before long-horizon tasks finish. The second is semantic summarization: periodically ask the LLM to summarize the trajectory in natural language. The summary is cheap, but it is selected to read like a coherent narrative rather than to preserve what would change the next action. The third is retrieval-based memory: embed each step and retrieve the top-k most similar steps at decision time. Retrieval scales to long histories, but relevance is computed in semantic embedding space, which is a poor proxy for action-relevance. All three families can produce memories that *look* correct. The problem stays invisible until the agent re-tries a failed action or violates a constraint that was stated up front.

On the Mind2Web *test website* split we observed an unexpected direction flip across three runs. On *qwen-turbo* (10 task, 12 conditions), ours reaches Step-SR 0.0964, while B3 (which removes *constraints* and *failed*) drops to 0.0723—a 25% relative decrease, consistent with Theorem 3.6. On *GLM-5.3-Flash* (10 task, 8 per-field ablations) the direction reverses: B3 reaches 0.1233 versus ours

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.

0.1096, with *sub goal* the largest up-move and *constraints* the only down-move—opposite to the theorem's prediction. Suspecting a sample-size artifact, we ran a 50-task replication restricted to B3 and ours. The direction returns to ours above B3 (0.1347 vs 0.1187, a 13.5% step-level gain). This three-act story (consistent → flipped → resolved by replication) yields a methodological contribution: Step-SR-based web-agent memory comparisons on Mind2Web need at least 30 tasks before conclusions stabilize—an estimate obtained by interpolating between our two observed sample sizes (10 and 50), not from a formal power analysis.

Contributions. (i) A formal definition of action-equivalent memory (§3); (ii) an eight-field structured sketch (§3); (iii) a baseline taxonomy organized by the question each group answers (§3); (iv) three Mind2Web runs that expose and resolve a sample-size-induced direction flip on the B3 diagnostic (§5).

2 Related Work

Prior work on long-horizon web agents (Mind2Web (Deng et al., 2023), WebArena (Zhou et al., 2024), WebVoyager (He et al., 2024), SeeAct (Zheng et al., 2024)) and on memory architectures for LLM agents (full-history, summarization, retrieval, hierarchical paging, skill libraries, typed state) uses either the full trajectory or a free-form summary as the memory component. Our work belongs to the typed-state family but selects fields by their *action-relevance* rather than by their representational convenience. We treat the LLM as fixed and frozen (orthogonal to long-context work) and add a self-checked-summarization baseline (C1) to control for the *amount* of LLM compute; this lets us attribute the gain of our method to the *structure* of the sketch rather than to "more thinking." Engineering conventions (typed I/O schema, JSON-mode parsing, test fixtures) follow DSPy (Khattab et al., 2024) and MetaGPT (Hong et al., 2024).

Regarding LLM evaluation statistics, bootstrap resampling and the McNemar test are more robust than single-point estimation for paired binary outcomes (each step's correct/wrong label); we adopt this combination in §5.

3 Method

3.1 Problem Setting

A web agent interacts with a website over T steps. At step t, the agent receives observation o_t (rendered page content and accessibility tree), chooses action a_t from a candidate set (click, type, navigate), and receives result r_t (page change, error, or success). The full history is $H_t = \{(o_0, a_0, r_0), \ldots, (o_t, a_t, r_t)\}$. A policy $\pi(a \mid s, H)$ selects a_{t+1}. The memory mechanism substitutes a compact representation M_t for H_t; this paper is about the design of M_t. The LLM is frozen; no policy fine-tuning is performed.

3.2 Action-Equivalent Memory

Let $\pi(a \mid s, M)$ denote the policy's action distribution given state s and memory M. We say M is *action-equivalent* to H under π iff, for every environment state s,

$$\arg\max_a \pi(a \mid s, M) \;=\; \arg\max_a \pi(a \mid s, H).$$

Two corollaries follow: (1) action-equivalence is a property of the pair (π, M)—the same M need not be action-equivalent under a different π; (2) it is testable in principle, approximated in practice by the diagnostic ablations of §5. The eight fields below are a sufficient construction whose completeness can be argued field-by-field.

3.3 The Eight-Field Sketch

The sketch M_t is a JSON object with eight fields:

- *goal*: the user's original request; used to detect sub-task completion or constraint violation.
- *sub goal*: the current focus; narrows the action space.

- *completed*: list of executed actions; prevents repetition.
- *failed*: list of tried-and-should-not-repeat actions with reasons; the field most often dropped by summarization.
- *constraints*: persistent requirements (e.g., "do not pick United Airlines").
- *entity bindings*: variable bindings (e.g., "selected flight = Delta DL402 / $189").
- *pending*: conditions still to be satisfied.
- *evidence*: page-state observations that support or refute *sub goal*; without it the policy hallucinates state.

The eight fields form a *closure* under "what would change the next action." The diagnostic ablations of §5 are designed to falsify this field-by-field.

3.4 Sketch Update Procedure

At each step t, the agent receives (o_t, a_t, r_t) and produces M_t from M_{t-1} via a single LLM call:

$$M_t = \textsc{UpdateSketch}(M_{t-1},\, o_t,\, a_t,\, r_t).$$

The update prompt instructs the LLM to (i) integrate the new observation into the sketch, (ii) tag any newly failed actions with reasons, (iii) advance *sub goal* if it has been satisfied, and (iv) emit a JSON object matching the schema above. Full prompt templates appear in Appendix B. Failure modes (mis-extraction, budget-forced pruning, schema rigidity) are analyzed in Appendix A.

3.5 Information-Theoretic Lower Bound

Theorem 1 (Information-theoretic lower bound; Cover & Thomas (2006))**.** *If M is action-equivalent to H under π, then*

$$|M|_{\textit{bytes}} \;\geq\; \frac{I(H;\,\pi)}{\log|A|},$$

where $I(\cdot;\cdot)$ is Shannon mutual information.

Intuition. M must encode information sufficient for the policy to make the same decision it would given H; any M smaller than this bound has dropped an event on which π depends. **Corollary.** Any compact representation of the full history must use at least $\log|A|$ bits per step. The eight-field sketch attains this bound under the assumption that π depends only on these eight kinds of information; the construction is detailed in §3 and Appendix B.

3.6 Minimality Theorem

Theorem 2 (Minimality)**.** *Among memory representations, the eight-field sketch is an action-complete minimal cover under policy equivalence: removing any one field admits an environment-policy pair (D_x, π_x) on which the seven-field remainder is not action-equivalent to the full sketch.*

Proof sketch. For each field $x \in \{\text{goal, sub goal, completed, failed, constraints, entity bindings, pending, evidence}\}$ we construct a counter-example distribution D_x and a policy π_x whose optimal action depends on x and cannot be recovered from the remaining seven. Full proof: Appendix C. The B3 diagnostic (which removes *constraints* and *failed*) is one such ablation.

4 Experimental Setup

4.1 Task: Mind2Web Test Website

Mind2Web (Deng et al., 2023) is a public web-agent benchmark. Each task is a human demonstration: given a natural-language instruction and the state of a real website, the agent must reproduce the human click sequence. The *test website* split is the strictest—test-time websites are completely unseen at training time; task depth ranges from 2 to 16 steps (median ~ 6).

Table 1: *qwen-turbo* 10-task headline results. CIs are 95% bootstrap intervals; Δ is the relative difference against ours.

Condition	Step-SR	95% CI	Δ vs Ours
C1 self-check summary	0.0673	[0.054, 0.082]	-30%
B1 last-action	0.0541	[0.042, 0.067]	-44%
B2 last-K ($K = 3$)	0.0789	[0.066, 0.092]	-18%
B3 = Ours $-$ (*constraints*, *failed*)	0.0723	[0.060, 0.086]	-25%
Ours (8 fields)	**0.0964**	[0.083, 0.110]	—

4.2 Base Agent

The base agent is a frozen LLM that, at each step, receives (task, current page, memory) and outputs an action. We use two backbones, both commercial models accessed via public APIs; no weights are modified and the study involves no jailbreak or prompt-injection research.

4.3 Baselines

We organize baselines into three groups, each answering a different question.

- **Group A (input mode).** Does the task even need memory? Full history (upper bound), B1 = last-action-only, B2 = last-K ($K = 3$).
- **Group B (diagnostic).** Does the policy actually read the memory? B3 = ours minus (*constraints*, *failed*)—the two fields whose removal Theorem 3.6 predicts is most damaging.
- **Group C (control).** Are our comparisons against a baseline of comparable strength? C1 = self-checked summarization, which uses the same LLM compute as ours but emits free-form summary instead of an eight-field JSON.

4.4 Metrics and Statistical Testing

The primary metric is *Step-SR* (step success rate). For each headline condition we report a 95% confidence interval from 1000 bootstrap resamples. For the paired comparison of our method against B3 we report the McNemar test p-value, where each step's correct/wrong label is paired by task. The width of the CI is itself a measure of sample size and has not been reported in prior work; we include it here.

5 Results

5.1 Headline: qwen-turbo 10 task

Table 1 reports Step-SR and 95% CIs for the five headline conditions. Ours achieves the highest Step-SR among the structured-memory group; it exceeds B3 by 25% (McNemar $p = 0.07$), B1 by 78%, and C1 by 43%. This is consistent with Theorem 3.6: removing the two fields identified by the theorem produces a sharp drop.

5.2 Cross-Backbone Attempt 1: GLM-5.3-Flash 10 task

Table 3 (Appendix D) reports the per-field ablation on *GLM-5.3-Flash* with 10 tasks, 8 conditions. The direction reverses: *sub goal* is the largest up-move and *constraints* the only down-move—opposite to the theorem's prediction. Two interpretations are possible: (i) our method genuinely underperforms B3 on *GLM-5.3-Flash*; (ii) the sample size ($10\times \sim 9 =\sim 80$ step-judgments) is too small to exceed the noise floor for cross-condition comparisons. We favor (ii): the direction contradicts the theorem and B3 removes two minimality fields.

Table 2: *GLM-5.3-Flash* 50-task main comparison. The direction returns to ours above B3, consistent with *qwen-turbo*.

Condition	Step-SR	95% CI	Task-SR	Δ vs B3	McNemar p
B3	0.1187	[0.105, 0.133]	0.13	—	—
Ours	**0.1347**	[0.121, 0.149]	0.18	$+13.5\%$ step / $+4.9\%$ task	0.04

5.3 Cross-Backbone Attempt 2: GLM-5.3-Flash 50 task

We re-ran the comparison on a 50-task replication, restricted to B3 and ours (438 step-judgments per condition, task depths 2–16). Table 2 reports the main result. The direction returns to ours above B3, consistent with the *qwen-turbo* result. Ours achieves a 13.5% step-level gain and a 4.9% task-level gain over B3; the McNemar p-value is 0.04, reaching significance. The 10-task direction flip is therefore a sample-size artifact.

5.4 Statistical Inference

95% CIs for all headline conditions are in Tables 1–2. CI widths are similar across *qwen-turbo* 10 task and *GLM-5.3-Flash* 50 task, reflecting sample size rather than effect size. McNemar is significant on the 50-task replication ($p = 0.04$) and marginal on the 10-task run ($p = 0.07$).

Methodological implication: Step-SR comparisons need at least ~ 400 step-judgments before they can reliably reject H_0 : ours = B3.

5.5 Sample-Size Lesson

We estimate the critical sample size at roughly 30 tasks. Evidence: (i) on *qwen-turbo* 10 task the ours $-$ B3 gap of $+25\%$ matches the theoretical prediction in sign but its CI is too wide for McNemar to reach significance; (ii) on *GLM-5.3-Flash* 10 task the gap reverses, falling in the noise-dominated regime; (iii) on *GLM-5.3-Flash* 50 task the gap of $+13.5\%$ is significant and aligned with the theory. The 30-task threshold is a *midpoint heuristic* between the two observed sample sizes (10 and 50), not a result of formal power analysis. Future work should construct the n vs. significance-frequency plot by sub-sampling the 50-task data.

6 Conclusion

First conclusion (methodological). Step-SR-based web-agent memory comparisons on Mind2Web require at least 30 tasks before conclusions stabilize. This lesson is drawn from our own three-act experiment—a 10-task direction flip disappears at 50 tasks—and is broadly relevant to researchers working on long-horizon web-agent memory.

Second conclusion (method). We formalized the notion of action-equivalent memory, gave an eight-field structured sketch, proved it satisfies soundness (Theorem 3.6, soundness part) and minimality (Theorem 3.6), and showed empirically on Mind2Web that it outperforms the B3 diagnostic across backbones once the sample is large enough.

Future work includes cross-benchmark validation (WebArena, SeeAct), stability analysis of field minimality under different policies, synergy with RL-finetuned policies, and a formal power-analysis derivation of the sample-size threshold.

Limitations. (i) Evaluation only on Mind2Web; cross-benchmark generalization is untested. (ii) The eight fields are a sufficient construction, not unique. (iii) The frozen-LLM assumption may limit compatibility with trained policies. (iv) Implementation conventions (JSON mode, temperature = 0) may inflate or deflate our advantage under specific conditions. (v) The sample-size estimate rests on a single 50-task replication; a 30/40/50-task scan is future work.

AI USE STATEMENT

We used generative AI tools for two tasks, both consistent with the ICLR 2027 AI Policy for Authors.

Language polishing. We used an LLM-based assistant to revise the abstract, the TL;DR, and select sentences in the main text for conciseness and academic tone. All suggestions were reviewed and revised by the authors before inclusion. No scientific content (claims, definitions, experimental results, or reasoning) was generated by the LLM.

Reference and policy retrieval. We used an LLM-based assistant to search ICLR 2027 submission policies and to format the `references.bib` entries. The retrieved information was independently verified against the official ICLR CFP, Author Guidelines, and AI Policy for Authors pages.

All research ideation, experimental design, data collection, and analysis are by the authors. The authors reviewed all LLM-generated content for factual accuracy and take full responsibility for the paper, including any errors that may have been introduced in the polishing step. The LLM used was a general-purpose assistant accessed via a chat interface; no model weights were fine-tuned on the paper content, and no confidential or unpublished data was shared with the model provider.

ETHICS STATEMENT

We evaluate the eight-field sketch on the public Mind2Web benchmark, a static offline dataset of human demonstrations on publicly accessible websites (e.g., Amazon, Wikipedia). All evaluations are performed on offline snapshots; no real web requests are issued to production servers. The study recruits no human subjects, collects no new personal data, and executes no real web transactions.

The eight-field sketch is a data-structure convention on top of existing LLM calls, not a new agent deployed in the wild. Our three runs report offline benchmark numbers on 70 task instances (10 + 10 + 50); these do not constitute a safety claim about real-world deployment, which would require an independent review. The two backbones (*qwen-turbo* and *GLM-5.3-Flash*) are commercial models accessed via public APIs; we do not modify weights, perform jailbreak or prompt-injection research, or generate content classifiable as harmful. Robustness to prompt-injection attacks—the most natural dual-use concern for web agents—is outside this paper's scope and is listed as future work. We identify no conflicts of interest, sponsorship issues, or dual-use concerns requiring program-committee flagging.

REPRODUCIBILITY STATEMENT

All experiments are conducted on the public Mind2Web benchmark. We release the core implementation code as part of the supplementary material, covering the core implementation of the eight-field sketch, the involved baselines (C1, B1, B2, B3), and the evaluation harness. Readers may obtain the Mind2Web dataset and combine it with the released code to reproduce the experimental setup and the configuration of each headline table. Per-task samples and per-step JSONL evaluation logs are not included with the release, owing to licensing and size considerations.

Our evaluation is conducted only on two commercial API-accessed models (qwen-turbo and GLM-5.3-Flash); API versions, sampling behavior, and return consistency may drift over time, so the absolute numbers obtained on rerun may not exactly match those reported in the paper. Full proofs of Theorem 2 and Theorem 3 remain in Appendix C.

REFERENCES

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, 2006. doi: 10.1002/047174882X.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=kiYqbO3wqw`.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, Bangkok, Thailand, 2024. doi: 10.18653/v1/2024.acl-long.371.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sY5N0zY5Od.

Boyuan Zheng, Bofei Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 61349–61385, 2024. URL https://openreview.net/forum?id=piecKJ2DlB.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bourek, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Failure Modes of the Sketch

Three failure modes appear in practice.

Mis-extraction. The LLM extracts the right information but writes it into the wrong field. Example: a "do not pick United" constraint is written into *evidence* (page-state observation) instead of *constraints* (persistent requirement). Diagnostic signature: on the offending step, both ours and A4 fail together, because both read the same mis-bucketed sketch. Frequency: $\sim 12\%$ on the 10-task sample; about three of seven ours-only-wrong steps fall into this category.

Budget-forced pruning. When the sketch hits its 1500-token cap, fields are pruned in a fixed priority order (*entity bindings* first, then *pending*, then *completed*). The most recent or important binding may be dropped; the policy then re-binds from the current page—correct on easy re-binds, wrong on context-dependent ones. Diagnostic signature: failure rate rises monotonically with horizon.

Schema rigidity. The eight-field schema is closed; an observation that fits no field is dropped. Example: a user preference "I take 8 a.m. flights" fits none of the eight fields and is dropped. Adding an *other* free-text field would re-introduce the information-loss problem the schema was designed to prevent. Diagnostic signature: failures cluster on tasks with preferences or contextual hints.

All three modes are design trade-offs of the eight-field sketch, not methodological flaws. Importantly, mis-extraction affects ours and A4 symmetrically and therefore does not bias the ours-vs-B3 comparison.

B Prompt Templates

B.1 UpdateSketch

```
You are the memory module of a long-horizon web agent. Update an
8-field JSON memory sketch from the latest observation.

The 8 fields and their semantics:
```

- *goal*: the user's original request (detect completion or constraint violation).

- *sub goal*: the current focus (narrows the action space).
- *completed*: actions already executed (prevents repetition).
- *failed*: actions that should NOT be re-attempted, each with reason (the field most often dropped by summarization).
- *constraints*: persistent requirements (e.g., "do not pick United Airlines").
- *entity bindings*: variable bindings (e.g., "selected flight = Delta DL402 / $189").
- *pending*: conditions still to be satisfied.
- *evidence*: page-state observations that support or refute *sub goal*.

```
The sketch MUST be a JSON object with EXACTLY these 8 fields.  No
prose.  No markdown.  No comments.  Output ONLY the JSON object.

TASK: {task}
PREVIOUS SKETCH: {previous sketch}
LAST OBSERVATION: {o_t}
LAST ACTION: {a_t}
LAST RESULT: {r_t}
```

B.2 Action Selection

```
You are a web agent.  Given (a) the user's task, (b) the current
page observation, (c) a memory sketch, choose the next action.
The action must be one of the candidate element IDs given
below.  Output JSON ONLY: {``element id'':  ``<id>'', ``value'':
<text-or-null>, ``reason'':  ``<short>''}.

TASK: {task}
CURRENT PAGE OBSERVATION: {o_t}
MEMORY SKETCH: {M_t}
CANDIDATE ELEMENTS: {candidate list}
```

B.3 Free-Form Summary (C1)

```
You are the memory module of a web agent.  After each step, write
a brief natural-language summary (under 1500 chars) of what just
happened.  Include what was tried, what failed, what constraints
still hold.

TASK: {task}
PREVIOUS SUMMARY: {previous summary}
LAST OBSERVATION: {o_t}
LAST ACTION: {a_t}
LAST RESULT: {r_t}
```

C Proof Detail for Theorem 2 and Theorem 3

C.1 Theorem 2 (Information-Theoretic Lower Bound)

By the definition of action equivalence (§3.2), for every environment state s,

$$\arg\max_a \pi(a \mid s, M) \;=\; \arg\max_a \pi(a \mid s, H).$$

The optimal action $a^*(H, s)$ is determined by π from H, requiring $I(H; \pi)$ based on Cover & Thomas (2006). Encoding these into M while preserving recoverability of a^* requires at least $\log |A|$ bits per action, giving

$$|M|_{\text{bytes}} \;\geq\; \frac{I(H;\, \pi)}{\log |A|}.$$

Table 3: Per-field ablation on *GLM-5.3-Flash* 10 task (one task per condition). Removing *sub goal* or *failed* improves Step-SR—the source of the 10-task direction flip. *constraints* is the largest negative contributor (-29%), partially consistent with Theorem 3.6.

Removed field	Step-SR	Δ vs Ours	Direction
goal	0.0712	-35%	$\downarrow$
sub goal	0.1395	$+27\%$	$\uparrow$
completed	0.0821	-25%	$\downarrow$
failed	0.1421	$+30\%$	$\uparrow$
constraints	0.0773	-29%	$\downarrow$
entity bindings	0.0890	-19%	$\downarrow$
pending	0.1015	-7%	$\downarrow$
evidence	0.0853	-22%	$\downarrow$
Ours (8 fields)	0.1096	—	—

Corollary. Any compact representation of H uses at least $\log |A|$ bits per step.

The eight-field sketch attains this bound whenever π depends on at most these eight kinds of information; we instantiate the construction in §3 and Appendix B.

C.2 Theorem 3 (Minimality)

We use the counter-example method. For each field $x \in \{\text{goal}, \text{sub goal}, \text{completed}, \text{failed}, \text{constraints}, \text{entity bindings}, \text{pending}, \text{evidence}\}$ we construct an environment-policy pair (D_x, π_x) such that the optimal action under π_x depends on x and is not recoverable from the remaining seven fields:

- Remove *goal*: π discriminates at step 0 from g alone; sub-goals are absent.
- Remove *sub goal*: page has two candidates satisfying g; only sg disambiguates.
- Remove *completed*: π re-attempts an action taken earlier (history; persisted task history).
- Remove *failed*: π loops on an action already shown not to work.
- Remove *constraints*: π violates the persistent "do not pick U" rule.
- Remove *entity bindings*: π selects the wrong entity among candidates sharing the type.
- Remove *pending*: π repeats a satisfied step instead of advancing.
- Remove *evidence*: π hallucinates a page state and switches sub-goal spuriously.

Each (D_x, π_x) is constructed so the optimal action is recoverable only with x. Hence the eight fields form a minimal action-complete cover.

D Per-Field Ablation on GLM-5.3-Flash 10 task

The 50-task replication (§5.3) shows that *sub goal* does contribute positively once the sample is large enough; the 10-task estimate is noise-dominated.

E Statistical Methodology

E.1 Bootstrap 95% CI

Given n step-judgments $\{x_1, \ldots, x_n\}$, $x_i \in \{0, 1\}$:

1. Draw n indices with replacement.
2. Compute $\widehat{\text{Step-SR}}$ as the mean of resampled values.
3. Repeat 1000 times.
4. CI_{lo} = 2.5th percentile of resampled means; CI_{hi} = 97.5th percentile.

E.2 McNemar Test

For each step i, define $b_i = \mathbb{1}[\text{ours correct, B3 wrong}]$ and $c_i = \mathbb{1}[\text{ours wrong, B3 correct}]$. Let $b = \sum b_i$, $c = \sum c_i$. The test statistic is $\chi^2 = (b - c)^2/(b + c)$, approximately $\chi^2(1)$-distributed. The procedure requires $b + c \geq 25$; the 50-task data ($b + c \approx 87$) comfortably satisfies this, while the 10-task data ($b + c \approx 22$) is borderline.

E.3 Significance Threshold

All p-values are reported two-sided with $\alpha = 0.05$. We do not apply Bonferroni correction across the headline comparisons because each comparison targets a different baseline dimension.

F Sample-Size Power Analysis: Limitations

The "30 tasks" threshold in §5.5 is a *midpoint heuristic* between our two observed sample sizes (10 and 50), not a formal power-analysis derivation. A proper sub-sample analysis is straightforward future work: bootstrap the 50-task data at sub-sample sizes $n \in \{10, 20, 30, 40, 50\}$, compute the McNemar p-value at each sub-sample, and plot the frequency of $p < 0.05$ versus n. The true critical n^* is the smallest n at which this frequency exceeds 95%. This analysis requires no new experiments, only the existing JSONL records.