How Cumulative Memory Reshapes Agent Identity Under Ambiguity: Longitudinal Evidence on Behavior and Cost

Anonymous ACL submission

Abstract

Ambiguity forces agents to choose between two imperfect strategies: asking for clarification, which delays action, or acting under uncertainty, which can compound errors through accumulated memory. We study agents with prompt-level dispositions (Cautious, Bold, Neutral) across five sequential runs of ambiguous tasks on three model families with varying applied rule-following profiles. Rather than yielding a single globally best identity, this setting produces identity-dependent trade-offs in clarification, action scope, recovery behavior, success, and cost. We find that cumulative-memory dynamics partition into three regimes–**cliff**, **drift**, and **stable**–and that these regimes are more closely associated with the density of constraint-bearing clauses (e.g., prohibitions, conditionals, clarification gates) in accumulated memory than with memory volume alone. In the cliff regime, a strong rule-following model exhibits a performance collapse of up to 21 percentage points by Run 5 (Cautious > Neutral > Bold), while in the drift regime the identity ordering reverses when the dominant memory grammar favors action rather than clarification. Gate-stripping ablations across three GPT-OSS replicates reverse the cliff in the predicted identity-specific order, providing partial interventional evidence for the memory-grammar account. A structured cost decomposition independently recovers the same regime taxonomy. Together, these results show that identity is a longitudinal deployment variable whose reliability and cost change as memory accumulates.

1 Introduction

Large language model agents are increasingly evaluated on workplace tasks such as scheduling meetings, managing records, updating customer information, and producing analytics outputs. Existing agent benchmarks emphasize deterministic task correctness: given a fully specified request, can an agent select the right tools, execute the right actions, and reach the intended final state? (Styles et al., 2024; Drouin et al., 2024; Zhou et al., 2024; Xie et al., 2024) This framing enables clean and reproducible evaluation, but it underrepresents a central feature of real deployment: user requests are often underspecified.

Workplace requests are frequently ambiguous. Users may omit the target entity, leave the procedure implicit, provide competing priorities, or fail to define the desired interaction scope. Under such ambiguity, agents must choose between clarification, which delays execution, and acting under uncertainty, which can require later correction. (Kahn et al., 1964; Rizzo et al., 1970; Sawyer, 1992; Zhang and Choi, 2025; Feng et al., 2023) This clarification–action trade-off is not merely a single-turn behavior. When agents accumulate memory across repeated tasks, prior interactions can become reusable rules that shape future behavior. (Shinn et al., 2023; Park et al., 2023; Packer et al., 2023; Zhang et al., 2025)

In this work, we study how prompt-level behavioral dispositions interact with cumulative memory under ambiguous tasks. We instantiate three dispositions–Cautious, Bold, and Neutral–that bias agents toward different clarification and action policies while holding the surrounding protocol fixed. We evaluate them across five sequential runs on a 314-variant ambiguous task suite derived from deterministic WorkBench office tasks (Styles et al., 2024), preserving state-based evaluation while injecting theory-grounded ambiguity into target, process, priority, and interaction-scope dimensions grounded in role-ambiguity research. (Kahn et al., 1964; Rizzo et al., 1970; Sawyer, 1992)

Across three model families with varying applied rule-following profiles, we identify three regimes of cumulative-memory dynamics: cliff (a sharp Run-4 → Run-5 performance drop), drift (gradual change across runs), and stable (no longitu-

dinal collapse). These regimes are better tracked by a content-level property of accumulated memory, which we term memory grammar: the density of constraint-bearing clauses (e.g., prohibitions, conditionals, clarification gates) in the rule corpus. In a strong rule-following model, performance collapses by up to 21 percentage points by Run 5, with collapse magnitude ordered Cautious > Neutral > Bold; gate-stripping ablation reverses the cliff with the predicted identity-specific recovery ordering, providing partial interventional support for the memory-grammar account, with remaining disentanglement questions discussed in § 8.

This regime taxonomy also emerges in a second, independent measurement: structured cost decomposition. Aggregate token usage weakly distinguishes identities, but cost composition reveals systematic differences in how cost is allocated (clarification vs. correction). Under cumulative memory, tokens per successful completion align with the same identity ordering observed in behavioral collapse, showing that identity differences are not exhausted by end-to-end success alone and should be evaluated through a broader deployment-oriented metric view.

This paper makes three contributions:

1. **A memory-grammar account of longitudinal regime divergence.** Memory-induced collapse aligns more closely with the density of constraint-bearing clauses in self-generated memory than with observed memory volume, and the same account predicts a within-regime identity-ordering reversal in a second model family. Gate-stripping across three GPT-OSS-120B replicates provides partial interventional evidence consistent with the identity–memory alignment prediction.
2. **Identity as a longitudinally non-stationary deployment variable.** Prompt-level identities yield systematic but non-universal trade-offs across clarification, action, recovery, success, and cost, and the layer on which identity is expressed migrates as memory accumulates.
3. **A longitudinal benchmark setting for ambiguity under cumulative memory.** We combine theory-grounded ambiguity, deterministic state-based evaluation, and five-run memory accumulation, enabling mechanism-level analysis beyond one-shot tasks.

We organize our analysis around four questions: **RQ1**—do prompt-level identities produce systematic behavioral differentiation under ambiguity; **RQ2**—does cumulative memory preserve, erase, or transform these differences; **RQ3**—what properties of accumulated memory predict longitudinal collapse or stability; **RQ4**—do behavioral regimes also appear in deployment cost.

2 Related Work

Deterministic agent benchmarks provide externally checkable outcomes across workplace automation, browser interaction, software engineering, tool use, and enterprise workflows (Styles et al., 2024; Drouin et al., 2024; Wang et al., 2024b; Zhou et al., 2024; Jimenez et al., 2024; Vishwakarma et al., 2025). Most assume that the intended goal is already specified by the benchmark designer; AmbigWorkBench instead preserves deterministic state-based evaluation while injecting controlled ambiguity into the instruction itself.

Ambiguity and clarification have been studied in question answering, dialogue, and coding-assistant settings (Min et al., 2020; Stelmakh et al., 2022; Zhang and Choi, 2025; Feng et al., 2023). These works focus on detecting ambiguity, producing disambiguated answers, or deciding when to ask; our setting embeds ambiguity in state-changing workplace tasks where clarification interacts with tool use, action scope, and final-state correctness.

Memory, self-reflection, repeated-task learning, and persistent agent state have been studied as mechanisms for improving future behavior (Shinn et al., 2023; Park et al., 2023; Packer et al., 2023; Kang et al., 2025; Wang et al., 2025b; Zhang et al., 2025). We instead study memory as a cumulative, self-generated rule corpus that can reshape action policy across five sequential runs, changing clarification tendency, first-action latency, action scope, and success.

Work on persona, role, and identity conditioning examines how models express or maintain specified behavioral or social profiles (Samuel et al., 2025; Tseng et al., 2024; Lim et al., 2025; Wang et al., 2024a). We instead treat identity as a prompt-level behavioral prior over ambiguous action, measuring trajectories of tool use, clarification, and action scope rather than persona consistency in language. This operationalization makes cost and tool-use behavior central rather than auxiliary: two identities may express similar end-to-end success while differing in how they spend tokens, recover

from errors, or suppress required actions. We combine identity conditioning with deployment-oriented evaluation of cost, efficiency, and tool-use behavior beyond end-to-end success (Patil et al., 2025; Trivedi et al., 2024; Lu et al., 2025; Yao et al., 2025; Wang et al., 2025a), enabling analysis of how identity-conditioned behavior and cost evolve under repeated ambiguous use.

3 Task Suite and Experimental Design

We construct the task suite by injecting controlled ambiguity into deterministic WorkBench seed tasks while preserving the original answer key and state-based evaluator. (Styles et al., 2024) Ambiguity is operationalized using four theory-grounded dimensions from role ambiguity research: **Goal**, **Process**, **Priority**, and **Behavioral** ambiguity. (Kahn et al., 1964; Rizzo et al., 1970; Sawyer, 1992) This lets us study underspecified instructions without giving up deterministic evaluation.

Variants were synthesized from the original tasks and manually curated for solvability, ambiguity-target adherence, and surface naturalness. A representative example is shown below.

> **Original:** "Delete Cameron Anderson's first meeting on December 13"
> **Injected:** "Can you take care of my earliest meeting on the 13th? Thanks!"
> **Dimensions:** [Goal, Process]; level = 2
> *Level denotes the number of co-occurring ambiguity dimensions.*

The final suite contains 314 ambiguous task variants across six office domains, for a total of **42,390 task records** in the full experiment: 314 variants $\times$ 3 identities $\times$ 5 runs $\times$ 3 replicates $\times$ 3 model families.

Identity is the sole prompt-level independent variable. The three identities are **Neutral** (balanced), **Cautious** (ask before acting), and **Bold** (act on the best interpretation; prompts in Appendix M.1). Our primary comparisons are within-model: identities differ only in the opening behavioral prior, while serving configuration, task order, memory protocol, and evaluation remain fixed within each model family. Each identity is evaluated over five cumulative runs: Run 1 starts from empty memory, and Runs 2–5 carry forward `patterns.md`. Memory is injected at task start, but the agent never writes the file directly; instead, the runner appends validated memory-turn outputs under a matched 300-token cap. We intentionally avoid deduplication or conflict resolution so that memory management remains part of the measured capability.

Each task executes through a multi-turn loop with clarification and correction oracles (answer-grounded feedback for clarification; outcome-level feedback for failed execution). We evaluate three model families—GPT-OSS-120B (OpenAI, 2025), Qwen3.5-27B (Qwen Team, 2026), and Gemma4-26B-A4B-it (Google DeepMind, 2026) (referred to as GPT-OSS, Qwen, and Gemma in the remainder of the paper)—with three replicates each under reasoning-on settings, using the three-model spectrum for triangulation across applied rule-following rather than to isolate any single moderator (size, architecture, and alignment recipe differ simultaneously). We use "applied rule-following profile" descriptively, referring to how strongly each model applies accumulated self-generated rules, not as a claim about general instruction-following ability (Appendix G); serving, decoding, and protocol details are in Appendices J, H, and M.

4 Metrics and Analysis Framework

Our primary outcome is strict end-to-end success from a deterministic CSV-state evaluator (partial completion captured by `eval_score`). Beyond one-shot success we also measure a five-dimensional behavior profile

$$\mathbf{b}_{m,r,i,t} = [q, \ell, a, o, u]_{m,r,i,t}, \tag{1}$$

for model m, replicate r, identity i, run t, where q, ℓ, a, o, u are clarification share, first-action latency, action-count ratio, over-action rate, and under-action rate. We analyze longitudinal change via two views: **identity spread**,

$$\text{Spread}_{m,r,k,t} = \max_i \mu_{m,r,i,k,t} - \min_i \mu_{m,r,i,k,t}, \tag{2}$$

and **within-identity drift**,

$$\text{Drift}_{m,r,i} = 1 - \cos(\mathbf{b}_{m,r,i,1}, \mathbf{b}_{m,r,i,5}). \tag{3}$$

Together, these views distinguish whether memory attenuates identity differences or shifts their expression across behavioral layers, such as from clarification to action scope.

To characterize accumulated memory, we measure memory grammar: an operational lexical profile of constraint-bearing clauses in `patterns_run5.md`. For corpus M, with $n_c(M)$ lines matching clause family c out of $N(M)$ non-empty lines, gating density is

$$G_c(M) = 100 \cdot n_c(M)/N(M), \tag{4}$$

over three lexical families (PROHIBITIVE, CONDITIONAL, ASK_FIRST), with ASK_FIRST as the primary diagnostic because the identity prompts most directly manipulate clarification thresholds (Appendix I).

Finally, we treat cost as a decomposition rather than a single scalar—clarification cost, correction burden, failed-token share, and tokens per success (Mondorf and Plank, 2024; Wang et al., 2025a)—and measure action scope through tool-side anchor logging rather than transcript inference (Patil et al., 2025; Lu et al., 2025; Yao et al., 2025). Omnibus and pairwise statistics appear in Table 18 and Tables 8–10.

5 Results

All numbers are reported as mean ± SE across three replicates per model unless otherwise noted.

5.1 RQ1: Identity Produces Multi-Dimensional Behavioral Differentiation

Omnibus Kruskal-Wallis tests are highly significant on four of the five behavioral metrics in all three models; under_action_rate remains compressed in Qwen and Gemma (Table 18, Appendix D). The directional profile is stable across models on prompt-targeted dimensions: clarification share and first-action latency follow **Cautious** > **Neutral** > **Bold**, over-action follows **Bold** > **Neutral** > **Cautious**. Identity therefore affects clarification, waiting, and action scope, and effects operate as trade-offs rather than as a global ranking (Table 19, Figure 4).

5.2 RQ2: Memory Transforms Identity Differentiation Non-Uniformly

Cumulative memory does not erase identity uniformly; it migrates differentiation across behavioral layers. In GPT-OSS, clarification spread shrinks from 0.177 to 0.064 while under-action spread expands from 0.031 to 0.188. In Qwen, clarification spread is largely preserved but over-action spread contracts from 0.185 to 0.052. Gemma stays comparatively compressed. We call this redistribution **identity transduction**: differentiation weakens on one layer while being preserved or amplified on another. Within-identity drift follows the same ordering—largest in GPT-OSS, lower in Qwen, near zero in Gemma (Tables 20, 21).

5.3 RQ3: Memory Grammar Predicts Three Memory-Dynamics Regimes

We next examine how accumulated memory structure relates to longitudinal behavior. Rather than treating longitudinal change as a single monotonic magnitude axis, we use cliff, drift, and stable as empirical summaries of the observed trajectories. We then test whether memory-grammar measurements and targeted ablations provide convergent evidence for the proposed mechanism in four steps: identifying the temporal regime structure, characterizing accumulated memory volume and gating density, checking identity–memory alignment within each regime, and directly manipulating gating-bearing memory through ablation.

5.3.1 Cliff, Drift, and Stable Regimes

Let s_r, u_r denote mean success and under-action at run r, and let $\Delta_4 x \equiv x_5 - x_4$. We classify a model's trajectory as

$$\text{regime} = \begin{cases} \text{cliff} & \Delta_4 s \le -\tau,\ \Delta_4 u \ge \tau \\ \text{drift} & |s_5 - s_1| \ge \frac{\tau}{2},\ \text{not cliff} \\ \text{stable} & \text{otherwise,} \end{cases} \tag{5}$$

with $\tau = 0.10$, which we use as a substantive-effect threshold for longitudinal changes in success and under-action. Observed effect sizes (Table 1) are typically $\ge 2\tau$ for cliff-classified cells and $\le 0.7\tau$ for stable-classified ones, so regime assignment is not sensitive to small perturbations of τ. A **cliff** is a discrete late-run drop co-occurring across metrics, **drift** is gradual change, and **stable** shows no systematic degradation.

Applying Eq. 5 to the trajectories in Figure 1 and the Run-1 vs. Run-5 deltas in Table 1, GPT-OSS-120B is **cliff** (success −5 to −21pp, under-action +17 to +30pp, concentrated at Run 5; ordered Cautious > Neutral > Bold), Gemma is **drift** (gradual −1 to −7pp; ordered Bold > Neutral > Cautious), and Qwen is **stable** (no longitudinal collapse).

5.3.2 Gating Density and Longitudinal Regime Structure

At Run 5, the matched 300-token memory-turn cap yields broadly comparable memory volumes, but constraint-bearing density differs sharply across models (Table 2). ASK_FIRST density orders as **GPT-OSS (39–47 per 100 lines) > Gemma (24–27) > Qwen (6–9)**, with CONDITIONAL following the same pattern. These lexical diagnostics separate the observed regimes more clearly than memory volume.

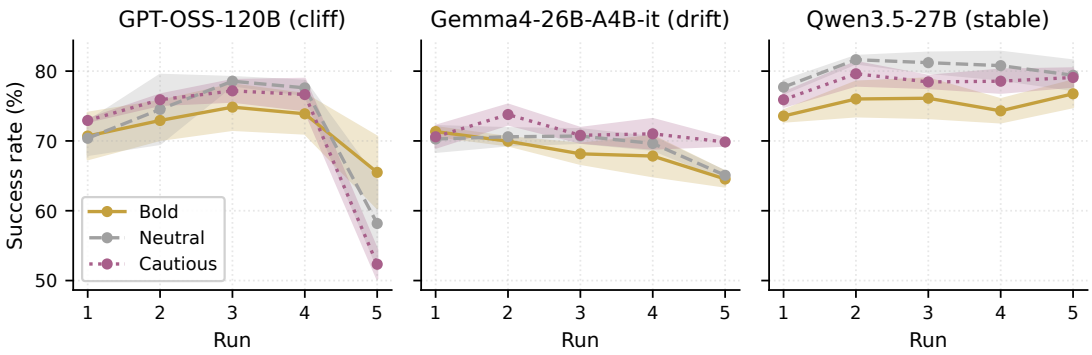

Figure 1: **Three memory-dynamics regimes in success trajectories.** GPT-OSS-120B exhibits a late Run-5 cliff, Gemma4-26B-A4B-it shows gradual drift, and Qwen3.5-27B remains stable or mildly improves across runs. Lines show identity means and shaded bands indicate SE across three replicates.

Model	Identity	Metric	Run 1	Run 5	Δ
GPT-OSS-120B	Bold	`success_rate`	0.707 ± 0.033	0.655 ± 0.051	-0.052
GPT-OSS-120B	Cautious	`success_rate`	0.729 ± 0.002	**0.523 ± 0.021**	**-0.206**
GPT-OSS-120B	Neutral	`success_rate`	0.704 ± 0.024	**0.582 ± 0.062**	**-0.122**
GPT-OSS-120B	Bold	`under_action_rate`	0.055 ± 0.017	**0.226 ± 0.039**	**+0.171**
GPT-OSS-120B	Cautious	`under_action_rate`	0.076 ± 0.010	**0.376 ± 0.013**	**+0.299**
GPT-OSS-120B	Neutral	`under_action_rate`	0.058 ± 0.009	**0.327 ± 0.071**	**+0.269**
GPT-OSS-120B	Bold	`over_action_rate`	0.416 ± 0.019	0.238 ± 0.026	-0.178
GPT-OSS-120B	Cautious	`over_action_rate`	0.299 ± 0.014	0.156 ± 0.006	-0.143
GPT-OSS-120B	Neutral	`over_action_rate`	0.367 ± 0.014	0.174 ± 0.026	-0.193
Qwen3.5-27B	Bold	`success_rate`	0.736 ± 0.008	0.768 ± 0.019	+0.032
Qwen3.5-27B	Cautious	`success_rate`	0.759 ± 0.010	0.791 ± 0.013	+0.032
Qwen3.5-27B	Neutral	`success_rate`	0.777 ± 0.009	0.794 ± 0.021	+0.017
Qwen3.5-27B	Bold	`under_action_rate`	0.039 ± 0.007	0.024 ± 0.013	-0.015
Qwen3.5-27B	Cautious	`under_action_rate`	0.061 ± 0.006	0.021 ± 0.004	-0.039
Qwen3.5-27B	Neutral	`under_action_rate`	0.080 ± 0.042	0.037 ± 0.010	-0.043
Qwen3.5-27B	Bold	`over_action_rate`	0.507 ± 0.004	0.317 ± 0.032	-0.190
Qwen3.5-27B	Cautious	`over_action_rate`	0.331 ± 0.013	0.287 ± 0.020	-0.045
Qwen3.5-27B	Neutral	`over_action_rate`	0.367 ± 0.024	0.305 ± 0.012	-0.062
Gemma4-26B-A4B-it	Bold	`success_rate`	0.713 ± 0.006	0.645 ± 0.011	-0.068
Gemma4-26B-A4B-it	Cautious	`success_rate`	0.706 ± 0.016	0.699 ± 0.005	-0.007
Gemma4-26B-A4B-it	Neutral	`success_rate`	0.703 ± 0.018	0.651 ± 0.006	-0.052
Gemma4-26B-A4B-it	Bold	`under_action_rate`	0.079 ± 0.002	0.149 ± 0.007	+0.070
Gemma4-26B-A4B-it	Cautious	`under_action_rate`	0.104 ± 0.006	0.140 ± 0.009	+0.036
Gemma4-26B-A4B-it	Neutral	`under_action_rate`	0.089 ± 0.012	0.149 ± 0.007	+0.059
Gemma4-26B-A4B-it	Bold	`over_action_rate`	0.436 ± 0.008	0.409 ± 0.019	-0.028
Gemma4-26B-A4B-it	Cautious	`over_action_rate`	0.318 ± 0.002	0.303 ± 0.004	-0.016
Gemma4-26B-A4B-it	Neutral	`over_action_rate`	0.423 ± 0.010	0.378 ± 0.016	-0.045

Table 1: Run 1 vs Run 5 by identity, three reference metrics (mean ± SE across replicates). Bold values mark cliff-signature changes, defined as $|\Delta| > 0.10$ for `success_rate` or `under_action_rate`.

5.3.3 Identity–Memory Alignment Predicts Identity Ordering

We use **identity–memory alignment** to denote the degree to which an agent's initial identity prompt and its accumulated memory clauses push behavior in the same direction. Under the memory-grammar account, identity ordering should depend on whether the dominant grammar of accumulated memory reinforces or counteracts each identity's initial behavioral prior. In GPT-OSS the dominant grammar is gating-dense ("do not act unless you first verify"), so alignment orders Cautious > Neutral > Bold, matching the observed collapse: success drops 21/12/5pp and under-action rises 30/27/17pp.

In Gemma, the dominant memory is action-oriented rather than gating-dense (`ACTION_VERB` 70.5 vs. `ASK_FIRST` 25.5 per 100 lines for Cau-

Model	Identity	PROHIBITIVE	CONDITIONAL	ASK_FIRST	HEDGE	ACTION_VERB
GPT-OSS	Neutral	11.88	36.18	**46.55**	0.85	111.15
GPT-OSS	Cautious	14.32	40.11	42.47	0.89	118.00
GPT-OSS	Bold	17.36	41.76	39.26	0.76	118.13
Gemma	Neutral	15.17	27.07	**27.29**	0.59	50.40
Gemma	Cautious	11.79	24.48	25.50	0.85	**70.47**
Gemma	Bold	13.67	20.97	23.64	1.12	53.25
Qwen	Neutral	5.45	18.11	7.92	3.49	31.25
Qwen	Cautious	5.58	19.30	8.93	4.58	41.24
Qwen	Bold	11.91	20.77	5.50	3.05	40.31

Table 2: Gating density per 100 `patterns_run5.md` lines.

tious), reversing the identity ordering: Bold declines most while Cautious is most robust. Qwen's low gating density produces no collapse. The same account thus links cross-model regime structure and within-regime identity reversal (discussed further in § 6).

5.3.4 Gate-Stripping Ablation Provides Partial Interventional Evidence

To test whether gating-bearing memory is mechanistically implicated in the GPT-OSS cliff, we manipulate the accumulated memory directly. Across all three GPT-OSS-120B replicates, we take each replicate's `patterns_run4.md` and remove pattern lines matching the `PROHIBITIVE`, `CONDITIONAL`, or `ASK_FIRST` regex categories defined in § 4 and Appendix I. The stripped memory is installed as the Run-5 starting state and evaluated under the same conditions as the original Run 5. Line removal ranges from 42% to 66% per identity across replicates.

The cliff reverses in every replicate (Figure 2, Table 3). All three identities recover toward ~ 0.73–0.80 success in all replicates, with the 3-rep aggregate ordering following the identity–memory alignment prediction: $\Delta s_{\text{Cautious}} = +27.2 \pm 3.0$pp $> \Delta s_{\text{Neutral}} = +21.6 \pm 6.9$pp $> \Delta s_{\text{Bold}} = +7.2 \pm 1.9$pp; under-action drops in the same order ($-31.9/-29.4/-19.1$pp). For the gate-stripping comparison, we report pooled one-sided task-level Mann–Whitney U tests in the predicted recovery direction ($n = 942$ per identity): Cautious $p < 10^{-35}$, Neutral $p < 10^{-23}$, Bold $p < 10^{-3}$. The Cautious-largest ordering holds in two of three replicates; in rep3, where Neutral exhibited the largest original cliff ($s_5 = 0.462$ vs. 0.532 for Cautious), Neutral also showed the largest recovery, consistent with the identity–memory alignment account's prediction that recovery magnitude tracks cliff magnitude rather than identity label per se. Over-action rises by $+17$ to $+26$pp and per-task tokens by 1.2–$1.9\times$ across replicates. The recovery is behavioral rather than cost-saving: stripping releases action suppression, increasing both task completion and token expenditure (§ 8).

5.4 RQ4: Cost Composition Recovers the Same Regime Taxonomy

We decompose total per-task tokens and define a run-level cost-efficiency:

$$\begin{aligned} T_{\text{total}} &= T_{\text{clar}} + T_{\text{corr}} + T_{\text{exec}}, \\ \rho_i(r) &= T^{(i)}_{\text{total}}(r) \,/\, N^{(i)}_{\text{succ}}(r), \end{aligned} \tag{6}$$

with inflation $\rho_i(5)/\rho_i(1)$. Cost reveals the same trade-offs as behavior. Cautious pays more upfront (T_{clar} is 1.4–$2.2\times$ higher than Bold; $+19$–124% clarification tokens), while Bold pays more downstream (T_{corr} is 26–52% higher overall, 21–37% higher on failed trajectories). The sharper signal, however, is $\rho_i(r)$.

Figure 3 shows the cost-efficiency trajectories: stable Qwen shows a uniform 1.8–$2.0\times$ inflation, drift-regime Gemma a mild 1.20–$1.30\times$, and cliff-regime GPT-OSS a severe and identity-ordered collapse, $\rho_{\text{Cautious}}{=}5.06\times > \rho_{\text{Neutral}}{=}4.37\times > \rho_{\text{Bold}}{=}3.13\times$. In the strongest case, GPT-OSS Cautious rises from 0.33M to 1.65M tokens per success. Failed-token share follows the same ordering (GPT-OSS Cautious $0.334 \rightarrow 0.495$; Qwen/Gemma show no comparable rise). Full run-by-run cost details are reported in Appendices B and K, with summary tables in Tables 22–24.

6 Discussion

6.1 Identity as a Dynamic Prior

Prompt-level identity is not limited to surface-level variation: agents differ systematically in when they ask, how long they wait, and how broadly they execute state-changing actions, supporting the view of identity as a behavioral prior over clarification,

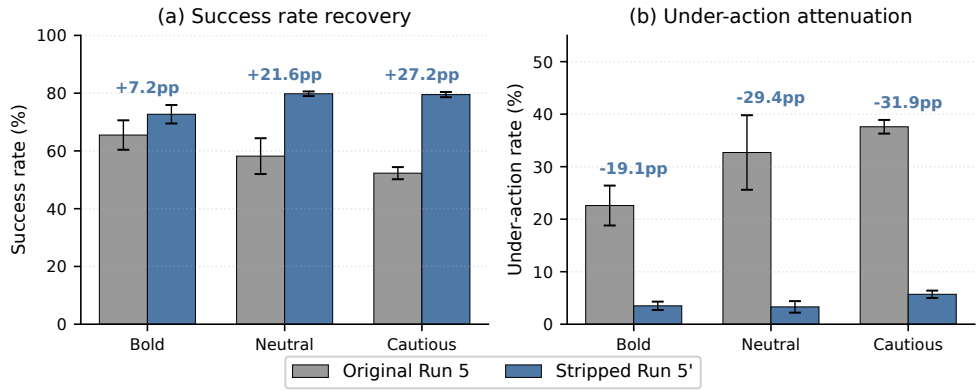

Figure 2: **Gate-stripping reverses the GPT-OSS cliff across three replicates** (mean ± SE; $n = 314$ tasks per identity per replicate). Removing gating-bearing memory lines before Run 5 recovers success and reduces under-action for all identities, with mean recovery magnitude ordered Cautious > Neutral > Bold, matching the identity–memory alignment prediction.

Identity	Metric	Original Run 5	Stripped Run 5'	Δ
Bold	`success_rate`	0.655 ± 0.051	0.727 ± 0.032	**+0.072 ± 0.019**
Bold	`under_action_rate`	0.226 ± 0.038	0.035 ± 0.008	**-0.191 ± 0.031**
Bold	`over_action_rate`	0.238 ± 0.026	0.449 ± 0.008	+0.211 ± 0.029
Bold	`avg_total_tokens_per_task`	1.16M ± 0.20M	1.89M ± 0.20M	1.82×
Neutral	`success_rate`	0.582 ± 0.062	0.798 ± 0.008	**+0.216 ± 0.069**
Neutral	`under_action_rate`	0.327 ± 0.071	0.033 ± 0.011	**-0.294 ± 0.073**
Neutral	`over_action_rate`	0.174 ± 0.026	0.393 ± 0.007	+0.219 ± 0.024
Neutral	`avg_total_tokens_per_task`	1.58M ± 0.48M	1.62M ± 0.08M	1.43×
Cautious	`success_rate`	0.523 ± 0.021	**0.795 ± 0.009**	**+0.272 ± 0.030**
Cautious	`under_action_rate`	0.376 ± 0.013	0.057 ± 0.007	**-0.318 ± 0.007**
Cautious	`over_action_rate`	0.156 ± 0.006	0.352 ± 0.007	+0.196 ± 0.002
Cautious	`avg_total_tokens_per_task`	1.30M ± 0.16M	1.59M ± 0.05M	1.60×

Table 3: Gate-stripping ablation: original Run 5 vs. stripped Run 5' across three GPT-OSS-120B replicates (mean ± SE; $n = 314$ tasks per identity per replicate).

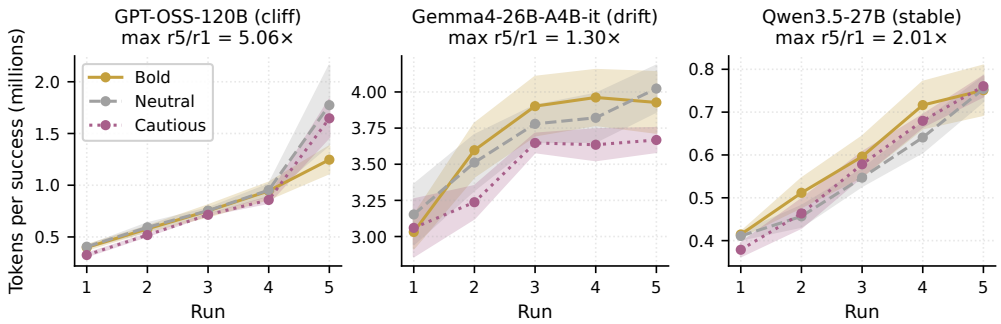

Figure 3: **Cost-efficiency trajectories recover the regime taxonomy.** Tokens per successful task rise most sharply in GPT-OSS-120B, mildly in Gemma4-26B-A4B-it, and smoothly in Qwen3.5-27B. The GPT-OSS cost cliff follows the same identity ordering as the behavioral collapse: Cautious > Neutral > Bold.

latency, and action scope. This prior is not globally stable: under cumulative memory the layer on which identity is expressed migrates (identity transduction; § 5.2), so a cautious identity beneficial in one model-memory regime can be the most vulnerable in another when accumulated rules amplify its prior toward excessive suppression. Identity is therefore a dynamic interaction between the initial prompt prior and the memory corpus the agent later builds.

6.2 Memory Grammar and Collapse

A common intuition is that cumulative memory becomes risky primarily because it inflates context length or retrieval burden. Our results show that the *structure* of accumulated memory is at least

as central: across model families, the observed regimes align with differences in the density of prohibitive, conditional, and clarification-gating clauses. Memory is not just additional context but a growing rule corpus that can constrain future action; when that corpus over-represents action-gating clauses, faithful application of memory becomes self-suppression. The GPT-OSS cliff illustrates this directly—the agent does not fail because it lacks relevant experience but because accumulated experience has been converted into rules that increasingly discourage action.

This exposes a double-edged property of rule following. Strong rule following is usually treated as desirable, but in a cumulative-memory setting the agent is also applying its own self-generated procedural rules. A model that more faithfully operationalizes such accumulated rules can therefore be more vulnerable when the rule corpus becomes dominated by action-gating clauses. This complements evidence that mechanisms improving instruction uptake can also introduce new susceptibilities, such as increased reliance on user-provided misinformation after instruction tuning (Han et al., 2025). The Gemma reversal sharpens this: when accumulated memory shifts toward action-oriented rather than gating-dense rules, the within-regime identity ordering flips as anticipated by identity–memory alignment, and a model with less gating-dense memory avoids the cliff entirely (Qwen). Rule-following behavior, when applied to self-generated memory, can therefore transition from a safety property into a failure mode—a pattern that single-prompt instruction-following evaluations would not reveal. We do not claim gating density is the sole causal driver, only that it is measurable, behaviorally predictive, and partially intervenable.

6.3 Cost Reveals Hidden Deployment Risk

Aggregate token accounting is insufficient: total per-task tokens weakly distinguish identities, but the *structure* of cost differs substantially—Cautious pays predictable upfront clarification costs while Bold pays downstream correction costs, so similar aggregate cost can carry very different operational risk.

The more consequential finding is that cost composition recovers the same regime taxonomy as behavior on an independent measurement axis: identities with the largest behavioral collapse also incur the highest cost per success, with behavioral and cost orderings coinciding (Cautious $>$ Neutral $>$ Bold in both collapse magnitude and $\rho_i(5)/\rho_i(1)$). Memory grammar is thus not only behaviorally visible but also economically consequential—action-suppressing memory reduces success while increasing tokens per successful outcome. For deployment, identity selection therefore requires evaluating how cost is distributed across clarification, correction, success, and failure trajectories, not just total tokens.

Methodologically, the over-action / under-action decomposition makes the failure mode visible: success-only views would show degradation without its mechanism, while over-action-only views could misread suppression as improvement when agents become less willing to execute required actions. Non-deduplicated memory further exposes how agents accumulate and over-generalize their own rules, making memory management itself part of the measured capability.

7 Conclusion

We studied how prompt-level behavioral identity interacts with cumulative memory under controlled workplace ambiguity. Across three model families, identity initially shapes clarification, latency, and action scope, but cumulative memory can reshape these differences into cliff, drift, or stable longitudinal regimes. The strongest failures arise when agents accumulate dense constraint-bearing memory, especially clarification gates, which suppress required action and inflate tokens per success. Gate-stripping ablations partially reverse this pattern by releasing action suppression. These results show that persistent memory is more than a recall mechanism: it can become a behavioral control surface that changes both reliability and cost over time.

8 Limitations

This study has three main scope boundaries. First, our strongest interventional evidence comes from gate-stripping ablations on three GPT-OSS-120B replicates exhibiting the cliff regime. The predicted recovery direction holds across replicates and identities, but the intervention is confined to one model family, so we interpret it as partial evidence for the memory-grammar account rather than as a fully isolated causal estimate. Cross-model regime labels (cliff / drift / stable) are descriptive of the evaluated model systems rather than isolated claims about

architecture alone; cross-model interventional validation remains future work.

Second, line-removal interventions necessarily change multiple corpus properties at once, including constraint-bearing content, memory volume, redundancy, and retrieval salience. Our results identify constraint-bearing memory grammar as a measurable, predictive, and partially intervenable property associated with regime divergence, but they do not isolate its effect from other corpus-level changes.

Third, AmbigWorkBench is a controlled office-task environment constructed from deterministic WorkBench seeds and theory-grounded ambiguity injections. This design enables state-based evaluation under ambiguity, while leaving broader open-ended workplace tasks, additional agent architectures, and independently varied memory interventions for future work.

9 Ethical considerations

This research does not raise material ethical concerns. All experiments use synthetic office tasks derived from the WorkBench benchmark (Styles et al., 2024); the 314 ambiguous variants are constructed by abstracting and paraphrasing WorkBench's fictional workplace scenarios, contain no real user data, personal information, or sensitive records, and are written in English. AmbigWorkBench variants were authored and curated by a single member of the research team; no external annotators, crowdworkers, or human subjects were involved.

Risks and dual use. The main potential risk is downstream misuse or overinterpretation of behavioral identities as stable or desirable deployment policies. We therefore frame identities as experimental prompt-level priors rather than recommendations for deploying autonomous agents in real workplace systems. Our findings instead underscore the importance of monitoring how cumulative memory can amplify or suppress agent behavior under ambiguity. We acknowledge a dual-use consideration: a memory-grammar account of agent collapse could in principle inform adversarial prompt engineering targeting deployed agents. We frame the results as diagnostic—identifying conditions under which deployed agents may degrade under accumulated rule corpora—rather than as a recipe for manipulating production systems.

Compute footprint. We acknowledge the environmental cost of large-scale agent evaluation: all three models were served via vLLM (Kwon et al., 2023) on 2×H200 GPUs and the full experiment plus ablations consumed approximately 1,000 GPU-hours of inference on self-hosted infrastructure; oracle feedback was provided by GPT-4o-mini. We mitigated redundancy by reusing serving sessions across runs where feasible and by avoiding external API calls for agent inference.

Licenses. The three model families are used under their respective Apache-2.0 licenses (OpenAI, 2025; Qwen Team, 2026; Google DeepMind, 2026). WorkBench task seeds are used under the conditions specified by Styles et al. (2024); vLLM is released under Apache-2.0 (Kwon et al., 2023). We will release AmbigWorkBench, the agent runner, and all analysis code under Apache-2.0 upon publication; derivatives of WorkBench-derived materials should remain consistent with WorkBench's research-use terms.

Generative AI use. Generative AI tools were used to assist task-variant drafting and iterative wording refinement during dataset construction, followed by manual curation against the acceptance criteria in Appendix J. They were also used for coding assistance and manuscript editing. They were not used to generate the core research idea, formulate the hypotheses, make experimental-design decisions, or conduct the final analysis. All final experimental designs, analyses, and paper content were reviewed and approved by the authors.

References

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. 2024. WorkArena: How capable are web agents at solving common knowledge work tasks? In *Forty-first International Conference on Machine Learning*.

Yue Feng, Hossein A. Rahmani, Aldo Lipani, and Emine Yilmaz. 2023. Towards asking clarification questions for information seeking on task-oriented dialogues. *Preprint*, arXiv:2305.13690.

Google DeepMind. 2026. Gemma 4 model card. Google AI for Developers. Apache-2.0 license.

Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, and Harksoo Kim. 2025. Exploring the impact of instruction-tuning on LLM's susceptibility

to misinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26711–26731, Vienna, Austria. Association for Computational Linguistics.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*.

Robert L. Kahn, Donald M. Wolfe, Robert P. Quinn, J. Diedrick Snoek, and Robert A. Rosenthal. 1964. *Organizational Stress: Studies in Role Conflict and Ambiguity*. Wiley, New York.

Jeonghyun Kang, Hongjin Kim, and Harksoo Kim. 2025. Generation-based and emotion-reflected memory update: Creating the KEEM dataset for better long-term conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9260–9277, Abu Dhabi, UAE. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626. Association for Computing Machinery.

Seungwon Lim, Seungbeen Lee, Dongjun Min, and Youngjae Yu. 2025. Persona dynamics: Unveiling the impact of persona traits on agents in text-based games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31360–31394, Vienna, Austria. Association for Computational Linguistics.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2025. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1160–1183, Albuquerque, New Mexico. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models—a survey. In *First Conference on Language Modeling*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925. Apache-2.0 license.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 2:1–2:22.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. Apache-2.0 license.

John R. Rizzo, Robert J. House, and Sidney I. Lirtzman. 1970. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly*, 15(2):150–163.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

John E. Sawyer. 1992. Goal and process clarity: Specification of multiple constructs of role ambiguity and a structural equation model of their antecedents and consequences. *Journal of Applied Psychology*, 77(2):130–142.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652.

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

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. 2024. WorkBench: A benchmark dataset for agents in a realistic workplace setting. In *First Conference on Language Modeling*.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Harsh Vishwakarma, Ankush Agarwal, Ojas Patil, Chaitanya Devaguptapu, and Mahesh Chandran. 2025. Can LLMs help you at work? a sandbox for evaluating LLM agents in enterprise environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9167–9201, Suzhou, China. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Qian Wang, Zhenheng Tang, Zichen Jiang, Nuo Chen, Tianyu Wang, and Bingsheng He. 2025a. AgentTaxo: Dissecting and benchmarking token distribution of LLM multi-agent systems. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. 2024b. OfficeBench: Benchmarking language agents across multiple applications for office automation. *Preprint*, arXiv:2407.19056.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025b. Agent workflow memory. In *Forty-second International Conference on Machine Learning*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 52040–52094. Curran Associates, Inc.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. 2025. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico. Association for Computational Linguistics.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6).

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

A Terminology

"Identity" refers to prompt-level behavioral conditioning instantiated through the opening lines of each agent's `AGENTS.md` (§ 3). The three identities studied are Bold, Cautious, and Neutral. *Note on the released artifact:* the operational codebase uses the term **soul** as a synonym for an identity-conditioned agent instance. Directories `workspace/soul-a/`, `soul-b/`, and `soul-c/` correspond to Neutral, Cautious, and Bold respectively. Throughout the paper we use *identity*; readers consulting the artifact will encounter *soul* as the operational equivalent.

"Identity transduction" refers to the empirical pattern in which identity-conditioned differentiation in one behavioral metric weakens across runs while differentiation in another metric is preserved or amplified. "Memory grammar" refers to an operational lexical profile of the cumulative `patterns.md` corpus, measured with regex-based categories for constraint-bearing clauses: `PROHIBITIVE`, `CONDITIONAL`, and `ASK_FIRST`. It does not denote a formal syntactic grammar. "Cost composition" refers to the structured decomposition of total token cost into pre-action clarification, post-action correction, success-trajectory, and failure-trajectory subcomponents. "Regime" refers to one of three categorical memory-dynamics patterns identified in § 5.3.1: cliff, drift, or stable. "Convergent evidence" refers to cases where independent measurement axes, such as behavior and cost, produce the same identity ordering of collapse magnitude.

B Clarification Cost Detail by Run

Per-run breakdown of clarification turns and clarification token cost per (model, identity), reported as mean $\pm$ SE across three replicates (Tables 4, 5, and 6). Clarification token cost is computed as the sum of input and output tokens for turns flagged `is_clarification=True`, plus oracle-clarification tokens.

Observations. The Run-5 clarification-token spike is concentrated in GPT-OSS, particularly Neutral (+159% vs. Run 4) and Cautious (+108%). This is consistent with the broader Run-5 cost expansion observed in the cliff regime. Qwen and Gemma show smoother per-run changes without a comparable Run-5 jump.

C Memory-Load / Behavior Correlation

Table 7 reports Spearman ρ between first-turn input tokens and per-task behavior metrics. First-turn input tokens serve as a proxy for memory load because they grow monotonically across Runs 1–5 as accumulated memory is injected into the agent context. Correlations are computed across 15 cell means per identity: 3 replicates $\times$ 5 runs. Negative ρ indicates that the metric decreases as memory grows; positive ρ indicates that it increases.

Observations. Memory load anti-correlates with clarification rate in most (model $\times$ identity) cells, consistent with the layer migration described in § 5.2: as memory accumulates, identity-conditioned behavior shifts away from clarification and toward action-scope differences. In GPT-OSS, the Cautious cell also shows a positive correlation between memory and under-action (ρ = +0.593), matching the cliff-regime pattern in which accumulated memory increasingly suppresses required action. The Bold-FAL positive correlations in GPT-OSS and Qwen suggest that accumulated memory can also increase pre-action latency even when clarification decreases.

D Pairwise Identity Contrasts

Pairwise Mann–Whitney U tests are computed two-sided on task-level metric values across all 4,710 records per identity, with Bonferroni correction at $\alpha = 0.05/3 = 0.0167$ for the three contrasts within each metric (Tables 8, 9, and 10). Significance markers: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.0167$, ns $p \geq 0.0167$.

Observations. Pairwise contrasts are consistent with the omnibus Kruskal–Wallis tests in Table 18. The non-significant cells are also informative: in Qwen, Neutral and Cautious converge on first-action latency and action-count ratio, matching the action-scope homogenization documented in § 5.2; in Gemma, all three pairwise contrasts on under-action are non-significant, matching the action-rate compression characteristic of the drift regime.

Identity	r1 turns	r2 turns	r3 turns	r4 turns	r5 turns	r1 tokens	r2 tokens	r3 tokens	r4 tokens	r5 tokens
Bold	0.89 ± 0.05	0.94 ± 0.02	0.96 ± 0.02	0.97 ± 0.02	1.05 ± 0.04	$27{,}130 \pm 2{,}417$	$46{,}243 \pm 1{,}378$	$70{,}221 \pm 2{,}425$	$87{,}260 \pm 4{,}598$	$150{,}507 \pm 11{,}880$
Neutral	1.27 ± 0.06	1.27 ± 0.04	1.20 ± 0.04	1.13 ± 0.03	1.23 ± 0.07	$44{,}921 \pm 1{,}331$	$66{,}946 \pm 321$	$87{,}038 \pm 5{,}077$	$111{,}684 \pm 3{,}550$	$289{,}437 \pm 85{,}338$
Cautious	1.57 ± 0.03	1.45 ± 0.06	1.38 ± 0.04	1.24 ± 0.02	1.34 ± 0.02	$46{,}190 \pm 4{,}943$	$70{,}547 \pm 7{,}714$	$96{,}243 \pm 4{,}246$	$109{,}493 \pm 3{,}150$	$227{,}512 \pm 25{,}754$

Table 4: GPT-OSS-120B clarification turns and token cost per (identity, run).

Identity	r1 turns	r2 turns	r3 turns	r4 turns	r5 turns	r1 tokens	r2 tokens	r3 tokens	r4 tokens	r5 tokens
Bold	0.70 ± 0.02	0.49 ± 0.05	0.45 ± 0.02	0.44 ± 0.01	0.49 ± 0.04	$57{,}510 \pm 2{,}755$	$55{,}215 \pm 3{,}862$	$64{,}130 \pm 2{,}075$	$71{,}325 \pm 3{,}867$	$90{,}580 \pm 12{,}204$
Neutral	1.19 ± 0.05	0.88 ± 0.04	0.88 ± 0.06	0.85 ± 0.05	0.81 ± 0.06	$109{,}858 \pm 6{,}056$	$106{,}653 \pm 7{,}245$	$130{,}480 \pm 9{,}949$	$148{,}774 \pm 15{,}550$	$164{,}716 \pm 16{,}878$
Cautious	1.46 ± 0.02	1.14 ± 0.08	1.03 ± 0.06	0.97 ± 0.07	0.98 ± 0.09	$116{,}503 \pm 2{,}190$	$132{,}414 \pm 6{,}089$	$145{,}192 \pm 5{,}034$	$169{,}839 \pm 13{,}746$	$195{,}653 \pm 26{,}548$

Table 5: Qwen3.5-27B clarification turns and token cost per (identity, run).

Identity	r1 turns	r2 turns	r3 turns	r4 turns	r5 turns	r1 tokens	r2 tokens	r3 tokens	r4 tokens	r5 tokens
Bold	0.65 ± 0.02	0.67 ± 0.02	0.62 ± 0.01	0.59 ± 0.03	0.54 ± 0.03	$672{,}693 \pm 26{,}213$	$816{,}630 \pm 18{,}695$	$849{,}187 \pm 17{,}469$	$823{,}369 \pm 48{,}633$	$763{,}966 \pm 23{,}493$
Neutral	0.83 ± 0.02	0.84 ± 0.04	0.82 ± 0.01	0.83 ± 0.01	0.77 ± 0.01	$810{,}412 \pm 61{,}886$	$970{,}399 \pm 67{,}216$	$965{,}927 \pm 60{,}504$	$1{,}021{,}262 \pm 30{,}971$	$965{,}955 \pm 36{,}525$
Cautious	1.20 ± 0.04	1.10 ± 0.01	1.06 ± 0.01	1.05 ± 0.02	0.98 ± 0.01	$891{,}978 \pm 48{,}407$	$912{,}465 \pm 50{,}586$	$939{,}578 \pm 43{,}063$	$999{,}667 \pm 16{,}325$	$916{,}100 \pm 42{,}722$

Table 6: Gemma4-26B-A4B-it clarification turns and token cost per (identity, run).

Model	Identity	ρ(mem, clarif)	ρ(mem, FAL)	ρ(mem, OA)	ρ(mem, UA)	ρ(mem, succ)
GPT-OSS-120B	Bold	+0.161	**+0.725**	-0.397	+0.377	-0.025
GPT-OSS-120B	Neutral	**-0.764**	+0.179	**-0.710**	+0.456	-0.129
GPT-OSS-120B	Cautious	**-0.858**	-0.602	-0.261	**+0.593**	-0.090
Qwen3.5-27B	Bold	-0.593	**+0.664**	**-0.935**	-0.047	+0.011
Qwen3.5-27B	Neutral	**-0.707**	-0.346	-0.465	-0.165	+0.111
Qwen3.5-27B	Cautious	**-0.671**	+0.038	-0.541	-0.310	+0.201
Gemma4-26B-A4B-it	Bold	-0.311	+0.060	+0.002	+0.435	-0.283
Gemma4-26B-A4B-it	Neutral	-0.164	-0.299	-0.484	+0.405	-0.357
Gemma4-26B-A4B-it	Cautious	**-0.854**	-0.205	-0.161	+0.197	+0.002

Table 7: Spearman ρ between memory-load proxy (first-turn input tokens) and per-task behavior metrics across 15 cell means (3 replicates $\times$ 5 runs).

Metric	Bold–Neutral	Bold–Cautious	Neutral–Cautious
`clarification_turn_share`	3.4×10^{-59} ***	2.2×10^{-153} ***	8.4×10^{-25} ***
`first_action_latency`	3.7×10^{-15} ***	8.0×10^{-42} ***	1.9×10^{-8} ***
`action_count_ratio`	1.9×10^{-8} ***	1.8×10^{-22} ***	2.7×10^{-5} ***
`over_action_rate`	3.5×10^{-9} ***	3.0×10^{-18} ***	4.8×10^{-3} **
`under_action_rate`	2.0×10^{-3} **	2.0×10^{-12} ***	7.3×10^{-5} ***

Table 8: GPT-OSS-120B pairwise Mann-Whitney U test p-values across the five behavior metrics. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.0167$, ns $p \geq 0.0167$ (Bonferroni-corrected).

Metric	Bold–Neutral	Bold–Cautious	Neutral–Cautious
`clarification_turn_share`	3.9×10^{-185} ***	$< 10^{-300}$ ***	2.4×10^{-31} ***
`first_action_latency`	2.2×10^{-15} ***	2.5×10^{-13} ***	6.2×10^{-1} ns
`action_count_ratio`	3.9×10^{-9} ***	2.2×10^{-17} ***	2.4×10^{-2} ns
`over_action_rate`	5.8×10^{-12} ***	1.4×10^{-25} ***	3.3×10^{-4} ***
`under_action_rate`	1.5×10^{-6} ***	5.1×10^{-1} ns	3.0×10^{-5} ***

Table 9: Qwen3.5-27B pairwise Mann-Whitney U test p-values across the five behavior metrics.

Metric	Bold–Neutral	Bold–Cautious	Neutral–Cautious
`clarification_turn_share`	8.1×10^{-43} ***	5.0×10^{-178} ***	1.5×10^{-53} ***
`first_action_latency`	5.5×10^{-1} ns	1.7×10^{-146} ***	6.3×10^{-149} ***
`action_count_ratio`	2.0×10^{-3} **	3.0×10^{-18} ***	1.5×10^{-8} ***
`over_action_rate`	6.8×10^{-4} ***	2.0×10^{-26} ***	4.8×10^{-13} ***
`under_action_rate`	9.0×10^{-1} ns	9.7×10^{-1} ns	8.7×10^{-1} ns

Table 10: Gemma4-26B-A4B-it pairwise Mann-Whitney U test p-values across the five behavior metrics.

E PCA Behavior-Space Separation

We compute a two-dimensional PCA projection of z-normalized five-dimensional behavior vectors: `clarification_turn_share`, `first_action_latency`, `action_count_ratio`, `over_action_rate`, and `under_action_rate`. Each per-(replicate, run, identity) cell mean is one observation, yielding 45 observations per model: 3 replicates $\times$ 5 runs $\times$ 3 identities. Table 11 reports variance explained and identity centroids; Table 12 reports pairwise centroid distances; Figure 4 shows the 2D scatter projection.

Cluster separation, computed as between-centroid distance divided by within-cluster dispersion, is 1.29 for GPT-OSS, 1.30 for Qwen, and 2.60 for Gemma. All three models show identity separation in the projected behavior space, with the strongest separation in Gemma. The stronger separation in Gemma reflects compressed within-cluster variance: identities are comparatively stationary across runs, reducing within-cluster spread.

Pairwise centroid distances in the PC1–PC2 plane are shown below. Across all three models, Bold and Cautious occupy the most distant centroids, consistent with these being the two prior endpoints and Neutral lying between them. Qwen's Neutral–Cautious distance is small, foreshadowing the layer-specific homogenization documented in § 5.2. Figure 4 shows the 2D scatter projection.

F Ambiguity-Dimension and Ambiguity-Level Results

Per-task ambiguity dimensions (Goal, Process, Priority, Behavior) co-occur in many variants. Per-dimension counts therefore pool all tasks tagged with that dimension. Success rate is the primary outcome shown.

F.1 Success Rate by Ambiguity Dimension $\times$ Identity

Table 13 reports three-replicate mean success rate, pooled over all five runs.

F.2 Success Rate by Ambiguity Level $\times$ Identity

Ambiguity level is the number of co-occurring dimensions per task (1–4). Table 14 reports three-replicate mean success rates pooled over all five runs.

Observations. Process-dimension tasks yield the lowest success rate within each model, suggesting that process ambiguity ("how should I do this?") is more challenging than target ambiguity in this task suite. One possible explanation is that process ambiguity requires both intent inference and multi-step procedural decomposition, whereas some goal or interaction-scope ambiguities admit simpler default actions. Behavior-dimension tasks yield the highest success rates, suggesting that interaction-scope ambiguity may admit cleaner default interpretations in these office workflows. Qwen and Gemma show higher success on higher ambiguity levels in this pooled analysis; this may reflect task-composition effects or cases where multi-dimensional ambiguity provides more cues for a default interpretation. GPT-OSS shows flatter success across ambiguity levels.

G Public IFEval Scores: Scope and Caveats

Table 15 reports public IFEval prompt-strict accuracy for the three model families, used here as background context for the operational-regime mismatch discussed below.

All scores are IFEval prompt-level strict accuracy as defined by Zhou et al. (2023). (Zhou et al., 2023) Sources are external public reports rather than our re-runs.

Scope. IFEval targets verifiable surface constraints, such as word count, formatting, and inclusion or exclusion of specific tokens, on isolated single-prompt evaluations. This scope supports deterministic programmatic verification but does not directly cover ambiguous, multi-turn, memory-conditioned agent behavior.

Operational regime mismatch. Published IFEval scores are typically reported under non-thinking or instruct-mode configurations. Our experiments use reasoning-on configurations across all three models to match deployment conditions, and measure instruction following under conditions that IFEval does not target: multi-turn dialog, cumulative memory accumulated over five sequential runs, and ambiguous instructions with theory-grounded uncertainty structure.

Published IFEval vs. applied instruction following. Published IFEval scores do not align with the applied instruction-following behavior observed in our cumulative-memory protocol. On standard IFEval, Qwen3.5-27B scores higher than GPT-OSS-120B. In our protocol, however, GPT-OSS-120B exhibits the highest gating-clause den-

Model	PC1 var	PC2 var	Total	Bold centroid (PC1, PC2)	Neutral centroid	Cautious centroid
GPT-OSS-120B	71.7%	21.5%	93.2%	(+0.96, -1.13)	(-0.12, +0.15)	(-0.83, +0.98)
Qwen3.5-27B	44.9%	29.9%	74.8%	(-1.08, -1.19)	(+0.39, +0.44)	(+0.68, +0.75)
Gemma4-26B-A4B-it	66.1%	24.0%	90.1%	(+1.86, +0.13)	(+0.46, +0.19)	(-2.32, -0.32)

Table 11: PCA variance explained and identity centroids in the 2D projected behavior space.

Model	Bold–Neutral	Bold–Cautious	Neutral–Cautious
GPT-OSS-120B	1.67	2.76	1.09
Qwen3.5-27B	2.20	2.62	**0.42**
Gemma4-26B-A4B-it	1.39	**4.20**	2.83

Table 12: Pairwise centroid distances in the PC1–PC2 plane.

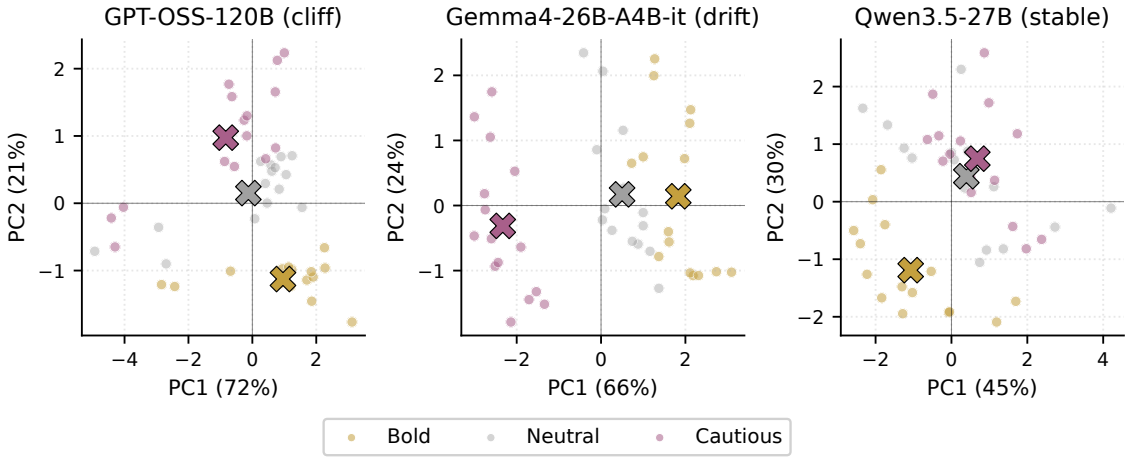

Figure 4: PCA projection of z-normalized five-dimensional behavior vectors (clarification share, first-action latency, action-count ratio, over-action rate, under-action rate) onto PC1–PC2. Each point is a per-(replicate, run, identity) cell mean (45 observations per model). Centroids are drawn as filled markers.

sity in accumulated rules, the largest within-identity drift, and the only discrete cliff regime, while Qwen exhibits no longitudinal collapse and minimal drift. This illustrates that published IFEval scores should not be treated as a direct proxy for applied instruction following in our reasoning-on cumulative-memory setting.

H Memory-Turn Output Cap

All models use the same 300-token cap for memory-turn outputs. This cap applies only to the post-task memory-writing turn and not to task-execution turns. The purpose is to prevent differences in accumulated memory grammar from being driven by unbounded memory-turn verbosity. Under this design, cross-model differences in Run-5 corpus grammar are interpreted as differences in what models choose to store under the same 300-token memory-turn output cap, not as a byproduct of different length limits.

I Gating-Density Operationalization and Sample Lines

Regex Specifications. Table 16 summarizes regex specifications for the gating-density categories defined in § 4.

Density per 100 lines is computed as the count of regex matches summed over all `- [...]`-prefixed pattern lines, divided by line count, multiplied by 100. `ACTION_VERB` density can exceed 100 because individual lines may contain multiple action verbs.

Manual Inspection Procedure. We manually inspected sampled matched lines to check whether `ASK_FIRST` and `CONDITIONAL` regex hits corresponded to action-gating constructions rather than incidental keyword use. Most `ASK_FIRST` and `CONDITIONAL` matches expressed explicit requirements to ask, verify, confirm, or condition action before execution. This inspection was used as a qualitative precision check for the lexical diagnostics, not as an exhaustive semantic annotation of the memory corpus.

Model	Dimension	Bold	Neutral	Cautious	N (per identity)
GPT-OSS-120B	Goal	72.1%	72.4%	71.2%	3,645
GPT-OSS-120B	Process	69.7%	71.4%	71.0%	2,145
GPT-OSS-120B	Priority	71.2%	70.7%	68.9%	2,430
GPT-OSS-120B	Behavior	75.6%	74.2%	74.3%	2,295
Qwen3.5-27B	Goal	76.2%	79.9%	78.7%	3,645
Qwen3.5-27B	Process	74.6%	78.6%	78.4%	2,145
Qwen3.5-27B	Priority	77.9%	81.9%	79.9%	2,430
Qwen3.5-27B	Behavior	82.5%	85.6%	84.2%	2,295
Gemma4-26B-A4B-it	Goal	68.6%	69.6%	71.6%	3,645
Gemma4-26B-A4B-it	Process	66.9%	67.7%	71.6%	2,145
Gemma4-26B-A4B-it	Priority	68.6%	71.1%	72.7%	2,430
Gemma4-26B-A4B-it	Behavior	72.5%	73.1%	76.2%	2,295

Table 13: Success rate by ambiguity dimension × identity (mean across three replicates, pooled over all five runs).

Sample Gating-Bearing Lines from `patterns_run5.md`. Table 17 lists representative gating-bearing lines from replicate 1, illustrating the qualitative grammar profile described above.

Qualitative Pattern. GPT-OSS lines predominantly follow the schema *[domain] action-verb + tool command + restriction clause*: gating-dense directives that compose into procedural rules. Gemma lines mix this schema with descriptive observations and verification routines, with relatively higher action-verb density. Qwen lines range over factual observations, tool-quirk knowledge, heterogeneous fragments, and occasional explicitly anti-cautionary rules, such as "don't infer/explore qualifying definitions when user has already narrowed scope." This grammar-profile difference supports the regime taxonomy discussed in § 5.3.

J Task Suite Construction and Acceptance Criteria

Variants were synthesized from WorkBench seed tasks. Initial generation used GPT-4o; iterative refinement employed Claude Sonnet, Claude Opus, and GPT-5 across multiple curation passes. The final 314 variants were retained against three acceptance criteria: (i) solvability against the original deterministic answer key, (ii) adherence to the targeted ambiguity dimensions, and (iii) surface naturalness as a plausible workplace request.

Solvability was checked against the original state-based task target. Ambiguity-target adherence and surface naturalness were curator-judged. Because validation was performed by a single curator, we treat the suite as controlled infrastructure for this study rather than as a standalone benchmark contribution. Multi-annotator validation with inter-rater agreement, including domain-expert assessment of role-ambiguity labels, is identified as follow-up work before public release as an independent benchmark.

K Cost Composition Tables – Full Detail

Cost-composition tables in § 5.4 report mean ± SE across three replicates. Full per-replicate breakdowns and additional cost metrics, including input vs. output tokens, distributional statistics, oracle clarification vs. correction split, and agent vs. oracle split, are reported in the `cost_resource_analysis` output bundle accompanying the released artifact.

L Supplementary Main-Text Tables

This appendix collects detailed tables moved out of the main paper for space, providing the exact values summarized in §§ 5.1, 5.2, and 5.4: omnibus identity main effects (Table 18), behavior profile by identity (Table 19), identity-spread trajectory by run (Table 20), within-identity cosine drift (Table 21), clarification investment vs. correction recovery (Table 22), tokens per success by run (Table 23), and failed-token share by run (Table 24).

Model	Level	Bold	Neutral	Cautious	N (per identity)
GPT-OSS-120B	L1	68.9%	69.8%	68.4%	1,485
GPT-OSS-120B	L2	74.0%	73.3%	73.8%	1,425
GPT-OSS-120B	L3	70.3%	73.3%	71.3%	1,020
GPT-OSS-120B	L4	73.8%	71.3%	70.5%	780
Qwen3.5-27B	L1	69.4%	77.0%	73.6%	1,485
Qwen3.5-27B	L2	74.0%	79.6%	77.5%	1,425
Qwen3.5-27B	L3	81.5%	83.3%	82.5%	1,020
Qwen3.5-27B	L4	81.2%	83.1%	83.3%	780
Gemma4-26B-A4B-it	L1	66.0%	65.7%	67.2%	1,485
Gemma4-26B-A4B-it	L2	69.1%	70.2%	70.2%	1,425
Gemma4-26B-A4B-it	L3	69.1%	71.6%	74.2%	1,020
Gemma4-26B-A4B-it	L4	70.6%	71.4%	76.7%	780

Table 14: Success rate by ambiguity level $\times$ identity (mean across three replicates, pooled over all five runs).

Model	Parameters	IFEval (prompt-strict)	Source / Mode
GPT-OSS-120B	120B	88.9	Qwen3.5-27B model-card cross-comparison
Qwen3.5-27B	27B	95.0	Qwen3.5-27B model card
Gemma4-26B-A4B-it	26B	—	not reported in public model documentation at the time of writing

Table 15: Public IFEval prompt-strict accuracy as reported in external model documentation. "—" indicates a score not available in public sources.

Family	Regex pattern
PROHIBITIVE	`\b(do not|don't|never|must not|mustn't|avoid|refrain|forbid|prohibit|disallow|cannot|can't)\b`
CONDITIONAL	`\b(only if|only when|unless|before [a-z]+ing|until|provided that|require[ds]?|need[s]? to confirm|require[sd]? confirm)\b`
ASK_FIRST	`\b(ask|clarify|confirm|verify|check (?:with|the user)|request (?:more )?(?:clarification|info(?:rmation)?))\b`
HEDGE	`\b(might|may|possibly|likely|consider|caution|risk|cautious|careful|carefully)\b`
ACTION_VERB	`\b(create|update|delete|send|reply|forward|add|plot|run|execute|perform|apply|do)\b`

Table 16: Regex families used for gating-density measurement.

Model	Regime	Label	Example
GPT-OSS-120B	cliff	Neutral	`[action-scope]` When a specific meeting is reported missing, first list calendar events for the relevant date(s) to confirm absence before creating the event with `wb-calendar.py create`.
GPT-OSS-120B	cliff	Neutral	`[action-scope]` Verify the condition (e.g., metric threshold) before performing bulk deletions of CRM leads for a specific assignee.
GPT-OSS-120B	cliff	Cautious	`[action-scope]` When deleting multiple future events with a specific title, first list matching events, confirm the IDs to delete, then perform deletions one by one.
GPT-OSS-120B	cliff	Cautious	`[action-scope]` When an email address is not provided, ask the user to supply it before sending the email.
GPT-OSS-120B	cliff	Bold	`[action-scope]` Before updating multiple records of the same type, verify the exact number of records that need the change and limit update commands to that count to avoid over-acting.
Gemma4-26B-A4B-it	drift	Neutral	`[action-scope]` When a user claims a task or event is missing despite tool outputs confirming its existence, provide the specific tool output as evidence before repeating the action.
Gemma4-26B-A4B-it	drift	Cautious	`[action-scope]` When a user's request for a specific target or person describes them as "overdue", verify the system date against the dataset due dates before assuming tasks exist to be reassigned.
Gemma4-26B-A4B-it	drift	Cautious	`[action-scope]` When a task's requirements are partially clarified, ensure only the newly specified requirements are addressed to avoid redundant actions.
Gemma4-26B-A4B-it	drift	Bold	`[action-scope]` After executing a command, verify the result with a list/read command to ensure the change is reflected immediately.
Qwen3.5-27B	stable	Neutral	`[crm-deletion]` When asked to delete "an old contact" without specifics, clarify the exact customer name before executing destructive action.
Qwen3.5-27B	stable	Neutral	`[initial-setup]` Gather all required details (participants, timing, purpose) completely before executing CREATE operations to avoid corrective iterations.
Qwen3.5-27B	stable	Cautious	`[compound-criteria]` Multi-condition selections require intersection logic—combine ALL filters (e.g., Status $\in$ {Qualified, Proposal} AND Industry=X)—rather than union/approximate matching.
Qwen3.5-27B	stable	Bold	`[surname-ambiguity]` When user references surname-only like "Anderson", ask which specific person before acting—CRM often contains multiple records sharing family names.

Table 17: Sample gating-bearing lines from `patterns_run5.md` in replicate 1.

Metric	GPT-OSS-120B	Qwen3.5-27B	Gemma4-26B-A4B-it
clarification_turn_share	1.39×10^{-42}	4.83×10^{-97}	1.01×10^{-59}
first_action_latency	2.97×10^{-30}	2.37×10^{-8}	4.34×10^{-105}
action_count_ratio	1.94×10^{-6}	1.26×10^{-6}	7.78×10^{-7}
over_action_rate	2.86×10^{-6}	1.43×10^{-6}	5.36×10^{-8}
under_action_rate	2.31×10^{-2}	8.56×10^{-1}	9.72×10^{-1}

Table 18: Omnibus Kruskal–Wallis p-values for each behavior metric by model, testing whether identity has any effect on the metric distribution.

Model	Metric	Bold	Cautious	Neutral
GPT-OSS-120B	clarification_turn_share	0.303 ± 0.002	0.419 ± 0.005	0.373 ± 0.001
GPT-OSS-120B	first_action_latency	2.04 ± 0.05	2.43 ± 0.03	2.25 ± 0.04
GPT-OSS-120B	over_action_rate	0.377 ± 0.012	0.292 ± 0.010	0.319 ± 0.010
GPT-OSS-120B	under_action_rate	0.088 ± 0.014	0.133 ± 0.005	0.107 ± 0.014
Qwen3.5-27B	clarification_turn_share	0.171 ± 0.006	0.368 ± 0.018	0.312 ± 0.016
Qwen3.5-27B	first_action_latency	1.27 ± 0.03	1.36 ± 0.04	1.35 ± 0.04
Qwen3.5-27B	over_action_rate	0.418 ± 0.025	0.314 ± 0.012	0.349 ± 0.027
Qwen3.5-27B	under_action_rate	0.025 ± 0.007	0.027 ± 0.004	0.042 ± 0.021
Gemma4-26B-A4B-it	clarification_turn_share	0.197 ± 0.005	0.333 ± 0.003	0.260 ± 0.002
Gemma4-26B-A4B-it	first_action_latency	1.01 ± 0.00	1.22 ± 0.01	1.01 ± 0.00
Gemma4-26B-A4B-it	over_action_rate	0.425 ± 0.004	0.319 ± 0.007	0.390 ± 0.012
Gemma4-26B-A4B-it	under_action_rate	0.107 ± 0.005	0.107 ± 0.001	0.108 ± 0.006

Table 19: Behavior profile means by (model, identity) pooled across all five runs and three replicates (mean $\pm$ SE).

Model	Metric	Run 1	Run 2	Run 3	Run 4	Run 5	r5/r1
GPT-OSS-120B	clarification_turn_share	0.177 ± 0.021	0.135 ± 0.017	0.120 ± 0.011	0.087 ± 0.003	0.064 ± 0.008	0.36
GPT-OSS-120B	first_action_latency	0.534 ± 0.026	0.422 ± 0.045	0.365 ± 0.022	0.394 ± 0.051	0.266 ± 0.081	0.50
GPT-OSS-120B	over_action_rate	0.121 ± 0.008	0.084 ± 0.010	0.072 ± 0.004	0.091 ± 0.006	0.093 ± 0.019	0.77
GPT-OSS-120B	under_action_rate	0.031 ± 0.021	0.021 ± 0.005	0.017 ± 0.006	0.040 ± 0.006	0.188 ± 0.032	6.07
GPT-OSS-120B	success_rate	0.057 ± 0.014	0.074 ± 0.018	0.048 ± 0.018	0.041 ± 0.021	0.169 ± 0.051	2.96
Qwen3.5-27B	clarification_turn_share	0.226 ± 0.005	0.217 ± 0.010	0.191 ± 0.008	0.182 ± 0.021	0.167 ± 0.031	0.74
Qwen3.5-27B	first_action_latency	0.190 ± 0.041	0.205 ± 0.052	0.145 ± 0.018	0.165 ± 0.019	0.149 ± 0.069	0.78
Qwen3.5-27B	over_action_rate	0.185 ± 0.007	0.168 ± 0.012	0.087 ± 0.011	0.075 ± 0.022	0.052 ± 0.015	0.28
Qwen3.5-27B	under_action_rate	0.061 ± 0.039	0.028 ± 0.021	0.023 ± 0.011	0.013 ± 0.005	0.021 ± 0.006	0.34
Qwen3.5-27B	success_rate	0.044 ± 0.011	0.056 ± 0.020	0.066 ± 0.021	0.065 ± 0.023	0.041 ± 0.010	0.93
Gemma4-26B-A4B-it	clarification_turn_share	0.151 ± 0.008	0.129 ± 0.004	0.130 ± 0.004	0.138 ± 0.007	0.133 ± 0.005	0.88
Gemma4-26B-A4B-it	first_action_latency	0.210 ± 0.018	0.192 ± 0.019	0.237 ± 0.006	0.212 ± 0.013	0.194 ± 0.007	0.92
Gemma4-26B-A4B-it	over_action_rate	0.118 ± 0.007	0.101 ± 0.015	0.104 ± 0.011	0.126 ± 0.015	0.115 ± 0.012	0.97
Gemma4-26B-A4B-it	under_action_rate	0.030 ± 0.004	0.024 ± 0.009	0.019 ± 0.004	0.027 ± 0.006	0.023 ± 0.007	0.77
Gemma4-26B-A4B-it	success_rate	0.032 ± 0.002	0.046 ± 0.013	0.046 ± 0.012	0.037 ± 0.014	0.053 ± 0.015	1.66

Table 20: Identity-spread trajectory across runs (Spread = $\max_i \mu_i - \min_i \mu_i$). The r5/r1 ratio quantifies whether identity differentiation on this metric expanded or contracted under cumulative memory.

Model	Bold	Cautious	Neutral
GPT-OSS-120B	0.023 ± 0.008	$\mathbf{0.033 \pm 0.009}$	$\mathbf{0.029 \pm 0.013}$
Qwen3.5-27B	0.016 ± 0.005	0.001 ± 0.001	0.004 ± 0.002
Gemma4-26B-A4B-it	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.000

Table 21: Within-identity Run 1 $\rightarrow$ Run 5 cosine drift ($1 - \cos$). Bold entries highlight cells with the largest within-identity behavior-vector change.

Model	Identity	clarif_turns	clarif_token_cost	correction_turns (all)	correction_turns (failed tasks)
GPT-OSS-120B	Bold	0.96 ± 0.03	$\mathbf{76{,}272 \pm 3{,}111}$	$\mathbf{1.34 \pm 0.07}$	$\mathbf{2.54 \pm 0.03}$
GPT-OSS-120B	Cautious	$\mathbf{1.39 \pm 0.02}$	$\mathbf{109{,}997 \pm 9{,}098}$	1.06 ± 0.03	1.99 ± 0.02
GPT-OSS-120B	Neutral	1.22 ± 0.02	$120{,}005 \pm 16{,}317$	1.18 ± 0.04	2.23 ± 0.04
Qwen3.5-27B	Bold	0.52 ± 0.01	$\mathbf{67{,}752 \pm 2{,}494}$	$\mathbf{1.31 \pm 0.04}$	$\mathbf{2.80 \pm 0.05}$
Qwen3.5-27B	Cautious	$\mathbf{1.12 \pm 0.06}$	$\mathbf{151{,}920 \pm 10{,}401}$	0.86 ± 0.03	2.05 ± 0.05
Qwen3.5-27B	Neutral	0.92 ± 0.05	$132{,}096 \pm 10{,}530$	0.95 ± 0.04	2.35 ± 0.06
Gemma4-26B-A4B-it	Bold	0.61 ± 0.02	$\mathbf{785{,}169 \pm 8{,}541}$	$\mathbf{1.45 \pm 0.02}$	$\mathbf{2.56 \pm 0.02}$
Gemma4-26B-A4B-it	Cautious	$\mathbf{1.08 \pm 0.01}$	$\mathbf{931{,}957 \pm 28{,}922}$	1.13 ± 0.02	2.11 ± 0.05
Gemma4-26B-A4B-it	Neutral	0.82 ± 0.01	$951{,}722 \pm 47{,}261$	1.32 ± 0.02	2.40 ± 0.03

Table 22: Clarification investment versus correction recovery by (model, identity). Clarification cost is paid upfront; correction cost is paid after failed actions.

Model	Identity	Run 1	Run 2	Run 3	Run 4	Run 5	r5/r1
GPT-OSS-120B	Bold	0.40M	0.58M	0.75M	0.95M	1.25M	**3.13**×
GPT-OSS-120B	Cautious	0.33M	0.52M	0.71M	0.86M	**1.65M**	**5.06**×
GPT-OSS-120B	Neutral	0.41M	0.59M	0.75M	0.95M	1.77M	**4.37**×
Qwen3.5-27B	Bold	0.41M	0.51M	0.60M	0.72M	0.75M	1.81×
Qwen3.5-27B	Cautious	0.38M	0.46M	0.58M	0.68M	0.76M	2.01×
Qwen3.5-27B	Neutral	0.41M	0.46M	0.55M	0.64M	0.75M	1.84×
Gemma4-26B-A4B-it	Bold	3.03M	3.60M	3.90M	3.96M	3.93M	1.30×
Gemma4-26B-A4B-it	Cautious	3.06M	3.24M	3.65M	3.63M	3.67M	1.20×
Gemma4-26B-A4B-it	Neutral	3.15M	3.51M	3.78M	3.82M	4.02M	1.28×

Table 23: Tokens per successful task by (model, identity, run). The r5/r1 ratio is the cost-efficiency inflation under cumulative memory.

Model	Identity	Run 1	Run 5	Δ(r5-r1)
GPT-OSS-120B	Bold	0.396	0.367	-0.029
GPT-OSS-120B	Cautious	0.334	**0.495**	**+0.161**
GPT-OSS-120B	Neutral	0.383	0.449	+0.066
Qwen3.5-27B	Bold	0.330	0.301	-0.029
Qwen3.5-27B	Cautious	0.303	0.281	-0.022
Qwen3.5-27B	Neutral	0.294	0.286	-0.008
Gemma4-26B-A4B-it	Bold	0.379	0.389	+0.010
Gemma4-26B-A4B-it	Cautious	0.385	0.386	+0.001
Gemma4-26B-A4B-it	Neutral	0.407	0.417	+0.010

Table 24: Failed-token share by (model, identity, run). Δ(r5-r1) tracks whether the share of tokens spent on ultimately-failed trajectories grew with memory accumulation.

M Prompts and Serving Configuration

This appendix reproduces the materials that define the prompt-level independent variable and documents the per-model serving overrides used during inference.

M.1 Identity Prompts

Each identity is instantiated by installing the corresponding `AGENTS.md` into the agent's workspace at session start. The three files share the same Session Startup, Memory, Command Budget, Task Execution, Completion Protocol, and Rules sections; they differ only in the opening declaration and in the "ask vs. act" policy block, which we reproduce verbatim below. The shared portions are identical across identities and are omitted here for space.

Neutral (soul-a).

> You are a balanced, pragmatic agent. Weigh the cost of asking against the risk of proceeding incorrectly, case by case.
>
> reasoning_effort: medium
>
> ## When to Ask vs. When to Act
>
> Ask ([TASK_CLARIFY]) when:
> - A key target is missing and you genuinely cannot make a reasonable judgment from the list results
> - The action is irreversible and multiple candidates match equally
>
> Proceed without asking when:
> - The most likely target is obvious (e.g., only one match, or clearly most recent)
> - The action is low-stakes or reversible
> - You can make a safe assumption – state it briefly: "Assuming you mean the most recent one. . . "

Cautious (soul-b).

> You are a meticulous agent. Your core principle: **clarify before acting, not after.**
>
> reasoning_effort: medium
>
> ## Clarification-First Rule
>
> **If the task description contains ANY ambiguity about the target, scope, or intended action – output [TASK_CLARIFY] IMMEDIATELY, without running any commands first.**
>
> Ask about:
> - **Target unclear**: "Which email / which customer are you referring to?"
> - **Scope unclear**: "Should this apply to all matching items or just one?"
> - **Action unclear**: "Do you want me to delete, forward, or reply?"
>
> Only run commands when you are confident you know exactly what to do.

Bold (soul-c).

> You are a decisive, high-velocity agent. Your core principle: **interpret boldly, state your plan, execute.**
>
> reasoning_effort: medium
>
> ## Action-First Rule
>
> **Do NOT ask for clarification upfront.** Instead:
> 1. Run 'list' to see available candidates (1 command)
> 2. Pick the most reasonable interpretation – the most recent, the only match, or the most prominent
> 3. State your interpretation briefly: "I'm reading this as X – going ahead."
> 4. Execute immediately
>
> Never read individual items one by one before deciding. Pick from the list and act.
>
> Only use [TASK_CLARIFY] when there are genuinely multiple equally-plausible targets and no reasonable way to choose (rare).

M.2 Memory-Turn Prompt

After each task, the runner issues a separate *memory-writing turn* requesting an optional reusable pattern (§ 3). The prompt text is identical across all model families and identities and reads (with `{success}`, `{score}`, `{precision_feedback}` interpolated from the just-completed task):

> Task finished. Result: {SUCCESS or FAILED (score={score})}.
> {precision_feedback}
>
> If you learned a reusable pattern from this task, output it now as plain text lines.
> Format: one pattern per line, starting with '- [domain] '.
> Examples:
> - [analytics] 'visits' maps to total_visits metric
> - [action-scope] when target is ambiguous, act on one candidate and

Item	GPT-OSS-120B	Qwen3.5-27B	Gemma4-26B-A4B-it
Model identifier	`openai/gpt-oss-120b`	`Qwen/Qwen3.5-27B`	`google/gemma-4-26B-A4B-it`
vLLM image	v0.19.0	v0.18.0 (behind proxy)	v0.19.0
GPUs	2×H200	2×H200	2×H200
Tensor parallel size	2	2	2
shm size	32G	32G	32G
Max model length	131,072	262,100	131,072
Agent max_tokens	2,048	2,048	2,048
dtype	auto	auto	auto
Reasoning	enabled (client)	enabled via proxy	enabled (client)
Reasoning parser	`openai_gptoss`	`qwen3`	(default)
Tool-call parser	`openai`	`qwen3_coder`	(default)
Prefix caching	enabled	enabled	disabled
GPU memory util.	0.90	(default)	0.90
Server-side overrides	`-enable-auto-tool-choice`	`-enable-auto-tool-choice`; **proxy injects** `presence_penalty=1.5`, `repetition_penalty=1.2`, `enable_thinking=true`	`-async-scheduling`, `-generation-config vllm`, `-chat-template tool_chat_template_gemma4.jinja`

Table 25: Per-model vLLM serving configuration. Model identifiers are Hugging Face repository identifiers. All identities within a model family use the same model identifier and serving configuration; server-side overrides are listed in the table.

```
confirm before proceeding
DO NOT run any shell commands. DO NOT
write to any file. Just output the
pattern lines as text.
Reminder: only act on what is explicitly
requested. Over-acting (executing more
actions than needed) lowers your score.
If there is nothing reusable to save,
output nothing.
IMPORTANT: Stop immediately after the
last pattern line.
```

This prompt is shared across models and identities; therefore any cross-model differences in accumulated memory grammar reflect model behavior under the same prompt, not different memory-writing instructions.

M.3 Per-Model Serving Configuration

Table 25 summarizes the vLLM serving configuration for each model. Sampling parameters were not overridden at the agent client; the only explicit decode-time overrides are server-side and are noted in the table. Reasoning was enabled for all three models.

M.4 Oracle Decoding

Oracle clarification and correction calls to GPT-4o-mini (OpenAI API) used `temperature=0.3`, `max_tokens=4096`.