CASM: Efficient LLM Medical Consultation via Selective Memory Materialization

Anonymous ACL submission

Abstract

LLM medical consultation agents must reason over growing multi-turn histories, but full-history prompting on every round spends latency and token budget even when a compact state already contains the clinically relevant evidence. CASM addresses this inefficiency by making memory exposure a controllable inference-time decision: each round starts from a structured compact view and materializes the full interaction trace only under uncertainty. Because the consultation backbone, patient simulation, and shared medical knowledge remain fixed, the evaluation isolates memory exposure rather than conflating it with a new diagnosis policy. On DXY test, CASM delivers a quality-aware efficiency point, reducing end-to-end latency by 15.1% while recovering 60.0% of the COMPACT–FULL accuracy gap; in a separate GMD transfer setting, it reaches higher accuracy than the matched FULL policy while cutting latency by 45.0%. These results position selective memory materialization as a practical and auditable control axis for efficient LLM agents.

1 Introduction

Large language models (LLMs) have made multi-turn medical consultation a realistic testbed for studying sequential information acquisition and differential diagnosis (Valizadeh and Parde, 2022; Tu et al., 2024; Li et al., 2024; Chen et al., 2024). In a typical trajectory, an agent receives self-reported symptoms, asks follow-up questions, updates disease beliefs, and finally outputs a diagnosis. The interaction history grows round by round, yet not every intermediate decision necessarily needs the full interaction trace.

Recent consultation frameworks have focused primarily on improving decision quality, for example through better inquiry planning, diagnosis calibration, or multi-agent coordination (Liu et al., 2022; Yan et al., 2023; Hou et al., 2023; Zhao et al., 2024; Hu et al., 2024; Jia et al., 2025). We ask a complementary systems question: *when should the system materialize the full history, and when is a compact state enough?* In many rounds, the next decision may depend mostly on confirmed positive symptoms, confirmed negative symptoms, asked/unasked status, remaining budget, and the current disease ranking. Repeatedly rematerializing the entire consultation history can therefore be wasteful, especially when prompt prefill dominates latency.

If consultation is treated only as a reasoning benchmark, this systems issue is easy to overlook. In serving stacks, repeated context materialization directly affects deployment cost. Long prompts consume tokens, increase prefill cost, reduce cache reuse, and make latency more sensitive to prompt layout and prefix stability (Kwon et al., 2023). The practical question is whether the agent should pay the cost of full history on every round.

Beyond consultation, recent agent-memory work has begun to treat memory as an active design space, studying efficient lifelong memory, evolving memory skills, and experience-driven skill improvement (Zhang et al., 2026; Liu et al., 2026; Liang et al., 2025; Yang et al., 2026; Ye et al., 2025; Yang et al., 2025; Yu et al., 2025). We take a narrower angle: given an existing consultation agent, can the agent avoid unnecessary full-history exposure through selective materialization at inference time?

This paper proposes a focused answer. We introduce CASM, an inference-time memory-exposure controller for LLM medical consultation agents. CASM maintains two dynamic views over the same consultation state: a *compact* view that serializes the current state into a stable structured representation, and a *full-history* view that exposes the complete interaction trace. Stable task instructions and disease-specific knowledge remain shared across both views. At each round, the agent

first reasons with the compact view and escalates to the full-history view only when uncertainty is sufficiently high.

CASM isolates memory exposure as an experimentally controlled intervention and turns it into a measurable policy variable. It can be inserted into an existing consultation pipeline without adding a learned router or a generative summarizer, while producing paired task–system measurements and an auditable routing trace. We study whether selective materialization creates a quality-aware efficiency region: lower latency than FULL inference, stronger quality than COMPACT inference, and enough routing selectivity to justify treating memory exposure as a policy decision rather than a fixed prompt template. The evaluation separates selection from measurement: DXY provides the primary held-out analysis, and GMD is analyzed as a separate transfer setting with its own dataset and hardware context.

Contributions.

1. **Problem formulation.** We recast full-history exposure in LLM medical consultation as a per-round resource-allocation policy.
2. **Method.** We propose CASM, a deterministic dual-view memory interface with uncertainty-triggered materialization of the full interaction trace.
3. **Evaluation.** We provide held-out matched comparisons, paired task–system summaries, and observed-policy diagnostics that expose both the efficiency gains and the residual cases where full history still matters.

2 Task Setup and Base Consultation Agent

We instantiate our study on top of a public multi-agent consultation framework (Jia et al., 2025). The base workflow decomposes consultation into four coordinated roles: a diagnosis module that estimates disease confidence, a policy module that proposes candidate inquiry actions, an inquiry module that selects the next symptom to ask about, and a patient simulator that returns the response. The system operates over a shared memory containing both static disease-related knowledge and a dynamic consultation trace.

This setting fits the memory-materialization question well. The dynamic consultation trace grows every round, but the agent also maintains structured state variables that already summarize much of the clinically relevant information: confirmed positive symptoms, confirmed negative symptoms, symptoms that have already been queried, turn budget, and the current disease-confidence distribution. The base agent already exposes two views of the same interaction: a *trajectory view* and a *state view*. CASM uses them as a selective interface.

This framework exposes the variables needed to study memory materialization: intermediate decisions are explicit, memory is updated round by round, and the task has bounded actions with clinically meaningful outcomes. We keep the diagnosis backbone, inquiry workflow, patient simulator, and dataset split fixed; the intervention is limited to how state is serialized, when compact reasoning escalates to full history, and how the prompt is organized. This focus isolates memory exposure from changes in the consultation policy.

3 Method

3.1 Design Principles

The controller follows four design principles. **P1: surgical intervention.** The method sits on top of an existing consultation agent without retraining its core modules. **P2: deterministic compact state.** If the compact view is itself generated by another unconstrained LLM, then any observed gain or degradation becomes difficult to attribute. **P3: transparent escalation.** The trigger is simple enough that readers can tell when it routes to full history and when it does not. **P4: measurable cost.** Each design choice has an observable footprint in latency, prompt length, prefill cost, or escalation behavior. These principles keep the intervention analyzable and separate it from training-heavy agent redesign.

3.2 Control Objective

CASM treats memory exposure as a per-round control problem. Let $v_t \in \{\texttt{compact}, \texttt{full}\}$ denote the memory view selected at round t, and let $c(v_t)$ denote its systems cost, measured through prompt length, prefill work, and latency. FULL fixes $v_t = \texttt{full}$ for all rounds, while COMPACT fixes $v_t = \texttt{compact}$.

Our goal is to choose an inspectable decision rule that reduces cumulative systems cost while

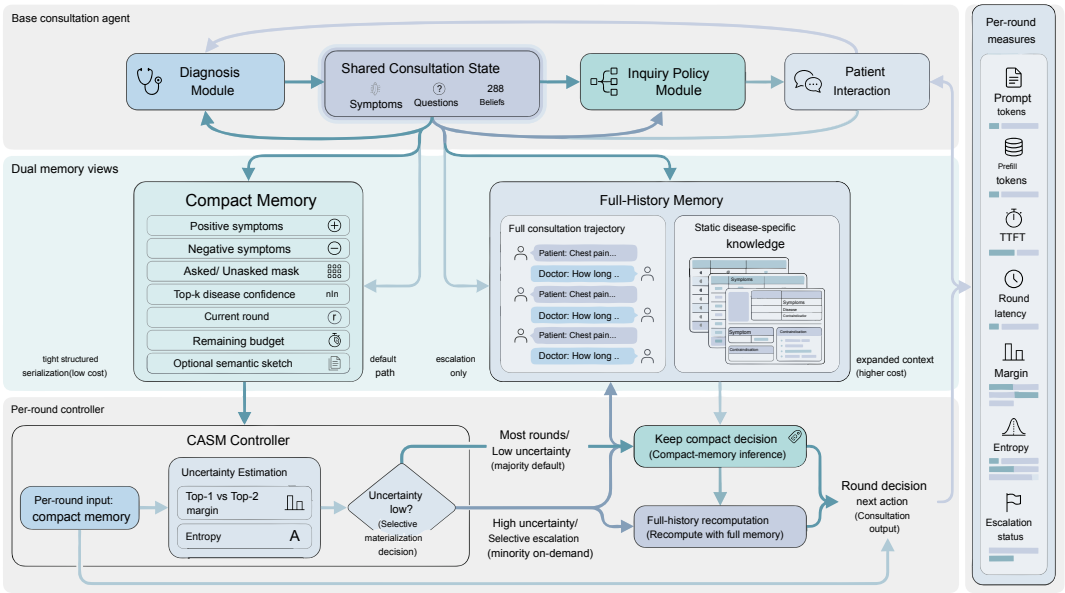

Figure 1: Overview of CASM. The base consultation pipeline is kept fixed while memory exposure becomes the inference-time decision variable. The controller switches between a compact structured state view and the full interaction trace, yielding both task and computational measurements for each round.

bounding final-quality loss:

$$\min_{\pi_{\text{view}}} \mathbb{E}\Big[\sum_{t=1}^{T} c(v_t)\Big] \quad \text{s.t.} \quad \Delta Q(\pi_{\text{view}}, \pi_{\text{full}}) \leq \epsilon, \tag{1}$$

where $Q(\cdot)$ denotes final task quality (e.g., diagnostic accuracy) and π_{full} is the FULL inference policy. CASM does not solve this optimization with a separate learned router. Instead, it uses uncertainty as a simple and inspectable proxy for whether the extra memory cost is warranted.

3.3 Two Memory Views

Full-history view. The full-history view materializes the complete consultation trace, including the original sequence of past inquiries and responses. This view is the most expressive dynamic representation, but also the most expensive to repeatedly serialize into prompts.

Compact memory. The compact view is a deterministic structured serialization of the consultation state. It includes:

- confirmed positive symptoms,
- confirmed negative symptoms,
- asked/unasked symptom mask,
- top-k disease confidence scores,
- current round index and remaining budget,
- an optional semantic sketch inherited from a prior state-augmentation prototype.

We study three codec variants: `slots`, `slots_topk`, and `slots_topk_d1`. The first uses only symptom-status slots; the second augments them with top-k disease beliefs; the third further adds the semantic sketch.

The serializer is deterministic after the base environment has produced its structured state. The patient simulation, symptom-status interpretation, and diagnostic distribution remain the same model-driven components used by the underlying consultation pipeline. This design avoids a second generative bottleneck whose own errors would be difficult to disentangle from the memory-control question. As a result, the matched comparisons isolate view selection more tightly than a design that also rewrites the state through a learned summarizer.

3.4 Uncertainty-Triggered Escalation

At round t, let p_t be the diagnosis distribution predicted under compact memory. CASM first runs the round with compact memory and computes uncertainty statistics from p_t. We consider the top-1/top-2 confidence margin

$$m_t = p_t^{(1)} - p_t^{(2)}, \tag{2}$$

and the entropy

$$H_t = -\sum_i p_t(i) \log p_t(i). \tag{3}$$

The controller escalates to full memory when uncertainty is high:

$$\texttt{use_full}_t = \mathbb{I}[m_t < \delta_m \ \vee \ H_t > \delta_H]. \tag{4}$$

We consider margin, entropy, and margin–entropy uncertainty triggers. The experiments below use the margin trigger, which gives a transparent scalar threshold for deciding when compact state is no longer confident enough. The view-selection interface yields three operating modes:

- **full**: always use full memory;
- **compact**: always use compact memory;
- **casm**: use compact memory by default and escalate only when triggered.

The trigger family is deliberately transparent. This choice makes the routing trace auditable: full-history materialization is tied to observed uncertainty signals instead of a fixed prompt requirement. The same simplicity also exposes a calibration boundary. If the compact-view distribution is overconfident, a margin trigger can fail to escalate when full history would help; if it is underconfident, routing can collapse toward the full-history baseline. We therefore report escalation density and paired residual errors together with accuracy, treating the uncertainty score as a routing signal rather than a calibrated probability. Figure 2 visualizes this decision process as a three-way comparison among FULL, COMPACT, and selective fallback.

3.5 Prompt Organization

CASM is motivated by token and prefill efficiency, so prompt organization matters. We place stable instructions and reusable policy scaffolding before dynamic memory fields whenever possible, using a shared ordering across memory modes. The same ordering is compatible with prefix-caching backends and paged serving engines (Kwon et al., 2023), but we do not include a separate fixed-layout ablation. We therefore make no independent layout-effect claim; the measured intervention is selective memory exposure under a fixed prompt organization.

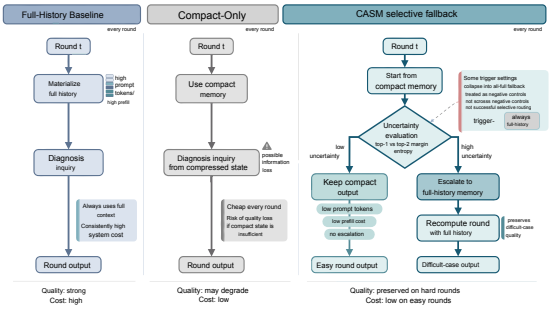

Figure 2: Round-level decision flow under three policies. FULL materializes the full interaction trace on every round, COMPACT never does, and CASM starts from the compact view before falling back to full history only when the uncertainty trigger fires.

3.6 Measurement Design

The evaluation measures both final outcomes and the path by which a case reaches them. At the case level, we measure diagnostic accuracy, number of consultation turns, end-to-end latency, and prompt usage. At the round level, we measure uncertainty, whether full history is materialized, and round latency. These measurements support mechanism-focused analyses. In particular, they let us analyze:

- whether escalation concentrates on later rounds,
- whether difficult cases trigger more fallback,
- whether observed residual errors align with routing density.

4 Experimental Setup

4.1 Benchmarks and Models

We use public medical consultation benchmarks centered on DXY, GMD, and CMD, following prior work on automatic consultation and multi-agent diagnosis (Chen et al., 2023b; Zhao et al., 2024; Chen et al., 2024; Jia et al., 2025). The experiments reported here use a 7B instruction-tuned backbone, which keeps threshold sweeps and per-round analysis tractable.

4.2 Baselines and Comparison Lines

We compare four inference policies:

- FULL: original full-history inference;
- COMPACT: compact memory only;
- CASM: standard selective-fallback configuration;
- $\text{CASM}_{\text{CONS}}$: a more conservative selective-fallback configuration.

More aggressive controller settings are treated as boundary cases unless evaluated under the same matched design as the main comparisons. The aggressive setting labeled `CASM_agg` showed that pushing escalation too low can trade away diagnostic quality, so it functions as a controller-selection boundary rather than a main result line. The main text otherwise avoids framework-specific naming because the contribution is the selective-materialization policy, not a renamed consultation baseline.

4.3 Metrics

We report final diagnostic accuracy (**Acc**) and average consultation turns (**Avg.n**), together with end-to-end latency, TTFT, prompt tokens, and escalation rate. For matched DXY comparisons, we also report paired case-level bootstrap intervals for accuracy and latency deltas. When backend instrumentation permits, we additionally inspect prefill-oriented measurements. Together these metrics describe the task–systems trade-off of each method.

4.4 Evaluation Design

Threshold selection and test reporting are separated. We tune thresholds on DXY development data, fix the selected policy, and evaluate it on DXY test as the primary held-out comparison. Additional datasets are reported as separate transfer settings rather than mixed into the DXY table. The DXY latency measurements come from a designated H200 environment; transfer latency measurements are interpreted only within their own matched dataset–hardware setting.

Two evaluation rules follow from this design. First, matched test comparisons determine the reported operating point; exploratory settings only reveal regime shape. Second, latency-facing numbers are interpreted within their hardware context. The DXY comparison uses one H200 environment, and transfer measurements are interpreted within their own hardware setting.

4.5 Hypotheses and Success Conditions

The paper is organized around three hypotheses. **H1**: COMPACT inference is a meaningful lower bound but not a sufficient policy. **H2**: selective fallback recovers part of that gap while still saving memory-related cost relative to FULL. **H3**: a successful controller remains genuinely selective rather than collapsing into almost-always-full routing. These hypotheses define the target operating region: lower systems cost than FULL inference, stronger quality than COMPACT inference, and a routing trace that can be audited at the case and round level. We use a dataset to support this operating-region claim only when matched FULL, COMPACT, and CASM evaluations are available.

5 Main Results

The main result is a quality-aware efficiency operating point for LLM consultation. Development data selects the controller, DXY test measures the primary trade-off, and codec variants show how the trade-off moves under alternative compact serializers. CASM achieves the intended regime: lower latency than FULL, stronger diagnostic quality than COMPACT, and an auditable selective-routing trace. We then report GMD separately as a transfer setting, keeping heterogeneous datasets and hardware contexts distinct.

Table 1 and Figure 3 give the primary DXY comparison. FULL reaches 0.728 accuracy with $43{,}206$ ms/case end-to-end latency, while COMPACT exposes the cost of removing full-history information: it is faster ($21{,}799$ ms/case) but drops to 0.583 accuracy. The selected CASM policy occupies the desired middle region, reaching 0.670 accuracy and $36{,}691$ ms/case, recovering 60.0% of the COMPACT–FULL accuracy gap while reducing end-to-end latency by 15.1% relative to FULL.

The conservative threshold ($m = 0.10$) escalates more often (0.710) but reaches lower accuracy (0.641), indicating that routing frequency alone does not determine the operating point. Codec choice and trigger calibration both affect where the selective policy lands, which is precisely why CASM exposes memory materialization as an analyzable control decision rather than a hidden prompt-design choice.

Table 2 gives the paired uncertainty estimate. The latency reduction is stable under case-level resampling, and the quality interval characterizes the DXY point as a controlled efficiency–quality trade-off. This paired view is important: the system gain is not merely a prompt-token reduction, but a case-level operating point with explicit uncertainty.

Prompt-token changes explain only part of the latency picture. On DXY, CASM reduces average prompt tokens by 6.2% and end-to-end latency by 15.1% relative to FULL. This mismatch is expected because case latency also depends on the

Method	N	Acc	ΔAcc	Gap rec.	Avg.n	Prompt tok.	E2E (ms)	ΔE2E	Esc.
Full-History Baseline	103	0.728	0.000	100.0%	4.3	3479	43206	0.0%	0.000
Compact-Only	103	0.583	-0.145	0.0%	4.3	2859	21799	-49.5%	0.000
CASM	103	0.670	-0.058	60.0%	4.5	3264	36691	-15.1%	0.649
CASM-Conservative	103	0.641	-0.087	40.0%	4.6	3316	40011	-7.4%	0.710

Table 1: DXY test comparison ($N = 103$ cases per policy). CASM recovers compact-only quality loss while reducing full-history latency. "Gap rec." measures recovery of the COMPACT–FULL accuracy gap; latency deltas are relative to FULL.

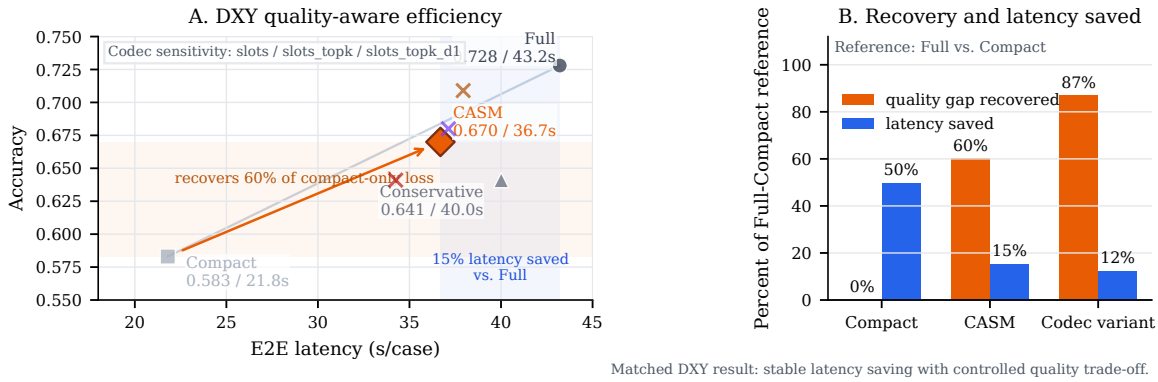

Figure 3: DXY quality-aware efficiency operating points. The selected CASM point recovers a substantial share of the compact-only quality loss while saving latency relative to FULL. The right panel decomposes each point into quality-gap recovery and latency saving; the codec variant is diagnostic rather than a replacement for the selected policy.

Comparison	ΔAcc (95% CI)	ΔE2E (95% CI)	Interpretation
Compact-Only	-0.146 [-0.233, -0.058]	-49.5% [-55.1%, -43.3%]	lower latency; below Full
CASM	-0.058 [-0.136, 0.019]	-15.1% [-23.5%, -5.7%]	latency lower; Acc CI crosses zero

Table 2: Paired bootstrap comparison against the DXY test FULL policy. Intervals use case-level resampling and show that the latency reduction is stable while the diagnostic-quality comparison remains a controlled trade-off.

Dataset	Device	Method	N	Acc	ΔAcc	E2E ms	ΔE2E
GMD	4090	Full	239/239	0.787	0.000	100986	0.0%
GMD	4090	Compact	239/239	0.724	-0.063	28464	-71.8%
GMD	4090	CASM	239/239	0.799	0.012	55511	-45.0%

Table 3: GMD transfer comparison. In this matched transfer setting, CASM improves the within-dataset accuracy anchor while substantially reducing end-to-end latency. ΔAcc and ΔE2E are computed against the within-dataset FULL policy.

reached turns, view-specific diagnostic refreshes, backend scheduling, and per-round prefill behavior. We therefore report token and latency measurements together, while reserving stronger layout or cache-causality claims for a fixed-layout control.

The codec variants in Figure 3 show that compact serialization changes the operating point. The `slots` codec reaches 0.709 accuracy with 37,934 ms/case latency and 0.697 escalation, while `slots_topk_d1` saves more latency but loses more quality. Because these variants were not the policy selected before test evaluation, they are reported as codec sensitivity rather than replacing the main comparison.

5.1 Transfer Setting: GMD

DXY is the split on which the controller is selected and analyzed in depth. To test whether the same memory-control idea can transfer beyond that setting, we additionally report GMD under its own matched dataset–hardware context. Latency deltas in this table are interpreted within the GMD setting, separately from the H200 DXY latency result. CMD requires complete matched FULL, COMPACT, and CASM evaluations before it can enter numerical comparison.

On GMD, CASM reaches 0.799 accuracy versus 0.787 for FULL, while reducing end-to-end latency by 45.0% within the matched RTX 4090 run. COMPACT is faster but loses quality (0.724), so the transfer setting strengthens the central claim: com-

pact state alone is insufficient, full history is expensive, and selective materialization can provide a high-value operating point for LLM consultation agents.

6 Mechanism Diagnostics

The mechanism diagnostics answer the questions that the main table cannot: which cases change, where residual errors concentrate, how difficulty interacts with routing, and whether the controller is actually selective. We report paired outcomes, disease-level residual buckets, FULL-margin difficulty buckets, and round-level fallback density, all derived from the same DXY test trajectories.

Table 4 and Figure 4 show the observed paired error profile. Compact-only inference leaves 20 full-only misses, while CASM reduces that bucket to 11, a 45.0% reduction relative to COMPACT, and retains 64 jointly correct cases. It also produces 5 CASM-only recoveries and 23 cases wrong under both views. The paired view separates recovered cases, residual full-history advantages, and errors that are shared by both memory policies.

Detailed taxonomy, disease buckets, difficulty buckets, and round-wise fallback densities are reported in Appendix A.5. They support the same mechanism claim without making the main text depend on appendix-only evidence: CASM recovers a meaningful share of compact-only misses, residual full-history advantages remain concentrated in identifiable pockets, and routing is selective rather than an all-full collapse.

Comparison	N	Both correct	Full only	Method only	Both wrong	Full-only red.
Compact-Only	103	55	20	5	23	0.0%
CASM	103	64	11	5	23	45.0%

Table 4: Paired DXY test outcomes against FULL. "Full only" means the baseline is correct while the compared method is wrong; "Method only" denotes the reverse direction. "Full-only red." measures reduction of full-only misses relative to COMPACT.

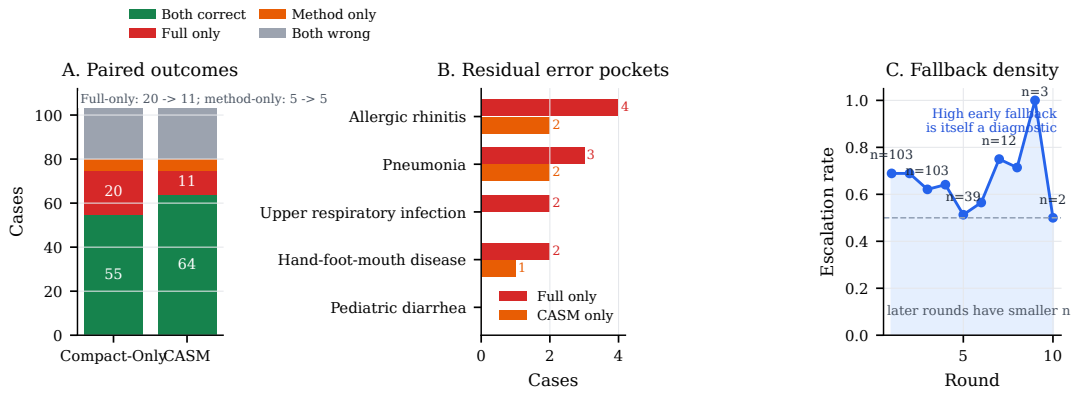

Figure 4: Mechanism diagnostics on DXY test: paired outcomes, residual disease pockets, and round-wise fallback density for the selected CASM policy.

7 Related Work

Automatic medical consultation has been studied with dialogue systems, reinforcement learning, generation-based diagnosis, and LLM-driven agents (Valizadeh and Parde, 2022; Chen et al., 2023b; Liu et al., 2022; Yan et al., 2023; Hou et al., 2023; Zhao et al., 2024; Tu et al., 2024; Li et al., 2024; Chen et al., 2024; Jia et al., 2025). These works mainly improve inquiry quality, diagnosis quality, or multi-agent coordination. CASM keeps this task substrate fixed and asks a different systems question: how much dynamic memory should an already capable consultation agent expose on each round?

CASM is also adjacent to work on agent memory, concise state descriptions, uncertainty-aware planning, selective prediction, and efficient LLM serving (Zhang et al., 2026; Liu et al., 2026; Liang et al., 2025; Nottingham et al., 2024; Hu et al., 2024; Chen et al., 2023a; Detommaso et al., 2024; Kumar et al., 2024; Xiong et al., 2024; Liu et al., 2024; Feng et al., 2024; Kwon et al., 2023). Its distinction is that uncertainty changes a concrete memory-exposure action, while the serving backend and consultation policy remain unchanged. Broader self-evolving skill and long-horizon memory systems (Yang et al., 2026; Ye et al., 2025; Yang et al., 2025; Yu et al., 2025) are therefore complementary rather than direct baselines: CASM is an inference-time control policy, not a lifelong memory learner.

8 Discussion

This work identifies a controllable interface inside an existing consultation agent: the memory-exposure decision itself. COMPACT defines a useful lower bound, selective fallback recovers compact-only misses on DXY, and the routing trace shows when the system pays for full-history materialization. Because CASM leaves the consultation policy, learned memory, inference backend, and shared disease knowledge unchanged, the matched comparisons narrow the intervention to routing over memory views. This attributable scope is the point: rather than changing retrieval content, summarization policy, and decoding context at once, CASM asks when compact state stops being enough.

9 Conclusion

This paper treats memory exposure in LLM medical consultation as a resource-allocation policy rather than a fixed prompt requirement. CASM operationalizes this view with a compact state codec and uncertainty-triggered full-history materialization while leaving the base consultation pipeline intact. On DXY test, the controller creates a measurable quality-aware efficiency regime between FULL and COMPACT inference: it saves latency, recovers much of the compact-only quality loss, and leaves an auditable routing trace. In the GMD transfer setting, the same controller reaches higher accuracy than the matched FULL policy while cutting latency.

The broader conclusion is methodological and systems-oriented. CASM makes LLM-agent memory selectable, measurable, and diagnosable at inference time, moving beyond compression ratios alone toward paired task–system outcomes. Across DXY and GMD, codec and threshold choices shape the quality–latency frontier.

10 Limitations

The DXY result should be read as quality-aware efficiency: CASM saves latency and recovers a large share of the compact-only quality loss, while residual full-history advantages remain visible. This boundary is productive rather than opaque. The paired failure tables, disease buckets, and difficulty buckets identify where the remaining gap concentrates, giving a concrete path for stronger trigger calibration and compact-state serialization.

Dataset coverage is governed by the same matched-evaluation rule used in the main paper. GMD already provides a complete transfer setting where the controller improves the within-setting accuracy anchor while reducing latency. CMD is excluded until FULL, COMPACT, and CASM are all available under the same matched protocol; once those rows are complete, they can enter the same table structure without changing the claim logic.

The main remaining systems uncertainty is variance estimation. The current comparisons keep latency-facing numbers within declared hardware settings, which supports the reported deltas, and repeated same-hardware runs would further tighten confidence intervals rather than change the mechanism being tested. Model-family transfer is also left to future work, but the interface is designed to be backbone-agnostic: CASM controls memory exposure after the base agent has produced structured state, so larger LLMs, calibrated uncertainty triggers, and fixed-layout cache studies can be added without redefining the method.

11 Ethics Statement

This paper studies automated medical consultation as a controlled research benchmark, not as a deployable clinical product. Real-world medical deployment would require clinician oversight, safety evaluation, data governance, and regulatory review. We also caution against interpreting systems efficiency improvements as inherently desirable in clinical contexts: faster inference matters only if diagnostic quality and safety remain acceptable.

References

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023a. Adaptation with self-evaluation to improve selective prediction in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023b. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. 2024. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10624–10641.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Zhenyu Hou, Yukuo Cen, Ziding Liu, Dongxue Wu, Baoyan Wang, Xuanhe Li, Lei Hong, and Jie Tang. 2023. Mtdiag: an effective multi-task framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14241–14248.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215.

Zhihao Jia, Mingyi Jia, Junwen Duan, and Jianxin Wang. 2025. DDO: Dual-decision optimization for LLM-based medical consultation via multi-agent collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26369–26386, Suzhou, China. Association for Computational Linguistics.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334.

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. *arXiv preprint arXiv:2309.06180*.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiafeng Liang, Hao Li, Chang Li, Jiaqi Zhou, Shixin Jiang, Zekun Wang, Changkai Ji, Zhihao Zhu, Runxuan Liu, Tao Ren, Jinlan Fu, and See-Kiong Ng. 2025. Ai meets brain: Memory systems from cognitive neuroscience to autonomous agents. *arXiv preprint arXiv:2512.23343*.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. 2024. Uncertainty calibration for tool-using language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16781–16805, Miami, Florida, USA. Association for Computational Linguistics.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Thang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. "my nose is running." "are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 4266–4272.

Kolby Nottingham, Yasaman Razeghi, Kyungmin Kim, Jb Lanier, Pierre Baldi, Roy Fox, and Sameer Singh. 2024. Selective perception: Learning concise state descriptions for language model actors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 327–341, Mexico City, Mexico. Association for Computational Linguistics.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Lian Yan, Yi Guan, Haotian Wang, Yi Lin, and Jingchi Jiang. 2023. Efficient evidence-based dialogue system for medical diagnosis. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3406–3413. IEEE.

Cheng Yang, Xuemeng Yang, Licheng Wen, Daocheng Fu, Jianbiao Mei, Rong Wu, Pinlong Cai, Yufan Shen, Nianchen Deng, Botian Shi, Yu Qiao, and Haifeng Li. 2025. Learning on the job: An experience-driven self-evolving agent for long-horizon tasks. *arXiv preprint arXiv:2510.08002*.

Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, Bo Zhang, and Liang He. 2026. Autoskill: Experience-driven lifelong learning via skill self-evolution. *arXiv preprint arXiv:2603.01145*.

Shicheng Ye, Chao Yu, Kaiqiang Ke, Chengdong Xu, and Yinqi Wei. 2025. H^2r: Hierarchical hindsight reflection for multi-task llm agents. *arXiv preprint arXiv:2509.12810*.

Simon Yu, Gang Li, Weiyan Shi, and Peng Qi. 2025. Polyskill: Learning generalizable skills through polymorphic abstraction. *arXiv preprint arXiv:2510.15863*.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026. Memskill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*.

Xuehan Zhao, Jiaqi Liu, Yao Zhang, Zhiwen Yu, and Bin Guo. 2024. Haiformer: Human-ai collaboration framework for disease diagnosis via doctor-enhanced transformer. In *ECAI 2024*, pages 1495–1502. IOS Press.

A Additional Experimental Details

A.1 Evaluation Setting

The results reported in Table 1 use a Qwen2.5-7B-Instruct backbone, the released consultation pipeline, and the CASM inference-time controller only. The DXY test comparison uses 103 cases and a `slots_topk_d1` codec selected on development data, with a margin trigger ($m = 0.08$), together with the released consultation limits of `max_turns=10` and `floor_turns=3`. The DXY latency table is measured on one designated NVIDIA H200 GPU. GMD is compared within its own matched RTX 4090 setting.

A.2 Latency Measurement Setting

Latency-facing measurements use warm-up runs before timing, keep each matched comparison on one declared hardware setting, and report prompt or prefill tokens whenever those measurements are available. The DXY comparison is measured on H200 and forms the primary latency comparison. The GMD transfer comparison is interpreted within its own RTX 4090 setting. Because repeated latency runs are outside the present evaluation, latency conclusions are stated as matched-run deltas rather than hardware-variance estimates.

A.3 Sanity Alignment with the Released Base Paper

Source	Acc	Use in this paper
Released base-paper DXY reference	82.7–87.5	Context only; split and reporting convention differ.
FULL in this work ($N = 103$)	0.728	Matched anchor for Table 1.
CASM in this work ($N = 103$)	0.670	Matched quality-aware efficiency point for Table 1.

Table 5: Sanity alignment between the released base-paper reference and the DXY results used here. The rows differ in split, seed, and reporting convention, so they contextualize the pipeline rather than define a cross-paper leaderboard claim.

A.4 Round-Level Analysis Definitions

The appendix makes the diagnostic quantities explicit. Let $R(x)$ denote the number of rounds in case x, and let $\mathbb{I}[\cdot]$ be the indicator function. We define the case-level escalation rate as

$$\mathrm{EscRate}(x) = \frac{1}{R(x)} \sum_{t=1}^{R(x)} \mathbb{I}[\texttt{escalated}_{x,t} = 1]. \tag{5}$$

At the dataset level, we report the average of $\mathrm{EscRate}(x)$ across cases.

The round-level traces include the selected memory view, uncertainty statistics, selected symptom, symptom status, prompt tokens, and round latency for every reached round. The mechanism tables use observed matched-policy trajectories rather than unobserved counterfactual deltas.

Difficulty buckets are defined using the full-history baseline as a reference model. Specifically, each case is assigned a difficulty score using the final-round margin under FULL. Cases are then partitioned into easy, medium, and hard buckets by quantile. This produces a stable analysis view in which selective fallback can be interpreted relative to a consistent baseline rather than to each method's self-reported confidence.

For round-level latency analysis, let $\ell_{x,t}$ denote the latency of round t in case x. The average round latency for a given turn index r is computed over all cases that reach that turn:

$$\mathrm{RoundLatency}(r) = \frac{\sum_x \mathbb{I}[R(x) \geq r] \cdot \ell_{x,r}}{\sum_x \mathbb{I}[R(x) \geq r]}. \tag{6}$$

This definition avoids biasing late-round measurements toward shorter cases that terminate early and makes the reported escalation-by-round plots easier to interpret.

A.5 Detailed DXY Mechanism Diagnostics

Outcome bucket	N	Share	Esc.	Operational reading
Compact sufficient	52	50.5%	0.500	Full, Compact, and CASM are correct
Useful recovery	12	11.7%	0.740	Full and CASM are correct; Compact is wrong
Residual Full advantage	11	10.7%	0.765	Full is correct; CASM is wrong
CASM-only recovery	5	4.9%	0.850	CASM is correct; Full is wrong
Shared residual	23	22.3%	0.809	Full and CASM are both wrong

Table 6: Residual outcome taxonomy for DXY test. Buckets are mutually exclusive and are derived from matched FULL, COMPACT, and CASM case outcomes.

Disease	N	Both correct	Full only	Method only	Both wrong
Allergic rhinitis	20	6	4	2	8
Pneumonia	20	4	3	2	11
Upper respiratory infection	23	21	2	0	0
Hand-foot-mouth disease	20	14	2	1	3
Pediatric diarrhea	20	19	0	0	1

Table 7: Disease-level residual paired outcomes for CASM on DXY test. Rows cover the DXY disease categories and should be read as residual error pockets, not as independent dataset-level claims.

Bucket	Criterion	N	Full	Compact	CASM	Esc.	Full only	CASM only
Hard	low Full margin	35	0.629	0.486	0.514	0.801	8	4
Medium	middle Full margin	34	0.676	0.500	0.676	0.777	1	1
Easy	high Full margin	34	0.882	0.765	0.824	0.344	2	0

Table 8: Difficulty-bucket diagnostics on DXY test. Buckets are formed by tertiles of the final-round FULL margin; lower margin means harder cases under the full-history reference.

B Prompt Layout and Memory Serialization

B.1 Compact Memory Fields

The compact memory serializer exposes:

- confirmed positive symptoms,
- confirmed negative symptoms,
- queried/unqueried symptom status,
- top-k disease confidence values,
- current turn index,
- remaining consultation budget,
- optional semantic sketch features.

B.2 Full Memory Fields

The full-history view exposes the original consultation trace. Stable disease-specific knowledge remains shared prompt context rather than a variable that changes across the compact and full-history views.

B.3 Prompt Ordering Strategy

The prompt layout follows a prefix-stable order:

1. stable system and policy instructions,
2. stable disease- or task-level scaffolding,
3. dynamic compact or full memory fields,
4. round-specific query target and output slot.

This ordering is held fixed across the reported comparisons. The paper does not attribute measured latency deltas to layout or cache reuse without a dedicated fixed-layout ablation.

Round	n	Esc. rate	Avg. margin	Avg. latency
1	103	0.689	0.0729	8409
2	103	0.689	0.0576	9029
3	103	0.621	0.0666	7957
4	64	0.641	0.0700	7514
5	39	0.513	0.0853	8238
6	23	0.565	0.0757	6823
7	12	0.750	0.0609	9487
8	7	0.714	0.0482	7319
9	3	1.000	0.0296	9301
10	2	0.500	0.0667	5747

Table 9: Round-wise escalation density for CASM on DXY test. Here n is the number of cases that reach the corresponding round.

B.4 Negative Controls

We treat two negative controls as bounds on the operating region. Any configuration with escalation rate $= 1.0$ and full-memory rate $= 1.0$ is an *all-full fallback control*. At the opposite extreme, a configuration that remains almost entirely compact while suffering a clear quality degradation is a *compact-insufficiency control*. The `CASM_agg` setting belongs to this safety-boundary family. Together these controls bound the target regime between indiscriminate fallback and insufficient memory exposure.

C Case Studies and Failure Analysis

C.1 Case-Study Evidence

The appendix reports case lists in place of free-form anecdotal examples. Each listed case carries the matched case id, disease label, number of turns, and end-to-end latency under FULL and CASM. This identifies residual pockets for qualitative inspection while keeping the analysis tied to matched case outcomes.

C.2 Residual Error Taxonomy

Table 6 organizes residual error analysis into five mutually exclusive categories:

- **compact sufficient**: FULL, COMPACT, and CASM are all correct;
- **useful recovery**: CASM recovers a COMPACT miss while matching FULL;

Case	Disease	CASM turns	Full turns	CASM E2E	Full E2E
test-env0-case19	Pneumonia	3	5	36859	50097
test-env0-case23	Pneumonia	4	3	50347	27301
test-env0-case40	Pneumonia	3	3	20735	36696
test-env0-case65	Hand-foot-mouth disease	3	4	28420	35574
test-env0-case72	Allergic rhinitis	4	3	31583	28216
test-env0-case73	Upper respiratory infection	6	5	58575	53166
test-env0-case74	Hand-foot-mouth disease	3	3	5992	31303
test-env0-case87	Allergic rhinitis	3	5	32624	40363

Table 10: Selected disagreement cases where FULL is correct and CASM is not on DXY test. Latencies are reported in milliseconds and can favor either method.

- **residual Full advantage**: FULL is correct while CASM remains wrong;
- **CASM-only recovery**: CASM is correct while FULL is wrong;
- **shared residual**: both FULL and CASM miss the target.

This taxonomy separates compact sufficiency, useful fallback, residual full-history value, and errors that remain outside the memory-exposure decision. The observed high escalation in residual buckets is a safety signal: improving the memory controller alone may not remove all remaining errors.

C.3 Replay Boundary

The current replay evidence consists of observed matched-policy trajectories: paired case outcomes, disease residuals, difficulty buckets, reached-round escalation, and listed full-only cases. Identical-state compact/full replay remains the missing analysis for round-level counterfactual margin claims.