ReMemorize: Adaptive RL-based Memory Management for Faithful Clinical Text Summarization

Anonymous ACL submission

Abstract

Large language models show strong potential for clinical summarization, yet factual inconsistency remains a major concern in long-context generation. Existing approaches largely rely on retrieval, verification, or post-hoc refinement, while leaving autoregressive memory dynamics fixed and allowing supporting evidence to be progressively overwritten during generation. We propose REMEMORIZE, a framework that introduces explicit control over information retention through a recurrent memory state and a learnable gating policy during decoding. The gating policy is optimized with reinforcement learning using sequence-level rewards for factual alignment, coherence, completeness, and hallucination suppression. Across two popular summarization benchmarks, `ReMemorize` improves factual consistency over strong baselines while maintaining competitive semantic and lexical quality. Our results suggest that explicit control over information retention is critical for faithful long-context clinical generation. Our code is publicy available at `https://anonymous.4open.science/r/ReMemorize-467B`.

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities in generative tasks, including clinical summarization, yet factual faithfulness remains a major challenge for their safe and reliable use in healthcare (Moeini et al., 2025; Nagar et al., 2025; Xie et al., 2023). Factual faithfulness is generally characterized by the preservation of source-supported information throughout generation (Ji et al., 2023; Honovich et al., 2022). Despite improvements in summarization quality, maintaining factual faithfulness remains a persistent challenge, particularly in long-context clinical settings where models must retain and accurately utilize distributed evidence. Recent benchmarks (Dong et al., 2024; Li, 2023) show promising gains from task-adaptation approaches, such as in-context learning (Brown et al., 2020), parameter-efficient fine-tuning methods (Dettmers et al., 2023), as well as post-hoc verification (Farquhar et al., 2024; Liu et al., 2025; Piya and Beheshti, 2026), retrieval augmentation (Borgeaud et al., 2022; Ram et al., 2023; Piya and Beheshti, 2025a), or iterative self-refinement (Madaan et al., 2023). However, these methods largely treat LLM memory dynamics as fixed. Recent work shows that LLMs often struggle to reliably utilize relevant information in long-context settings, exhibiting positional biases and degraded performance as context length increases (Liu et al., 2024). Memory in LLMs can refer to (i) the parametric memory stored in model weights, (ii) external memory accessed through retrieval or tool use, and (iii) working memory maintained through token-level hidden states during inference (Meng et al., 2022; Borgeaud et al., 2022; Schick et al., 2023).

While retrieval augmentation (Ram et al., 2023; Shi et al., 2024) and agent-based memory systems (Park et al., 2023; Packer et al., 2023) improve access to relevant information, they do not explicitly regulate how information evolves once generation begins. These observations suggest that faithfulness-related hallucinations in long-context generation arise not only from failures in retrieval or source compression, but also from limitations in adaptive memory regulation during decoding. In autoregressive transformers, working memory is implicitly governed by hidden-state transitions and attention caches whose update dynamics are fixed by the architecture (Khandelwal et al., 2018; Sun et al., 2021; Mallen et al., 2023). Consequently, models lack explicit control over which information is preserved, overwritten, or forgotten during decoding, contributing to factual drift in long-context generation. Targeting working memory allows for efficient governance of how contextual evidence is retained throughout autoregressive de-

coding.

In this study, we propose REMEMORIZE, a framework that introduces a learnable gating mechanism to regulate information retention throughout generation, unlike standard architectures where memory evolution is implicitly determined by fixed attention dynamics (Vaswani et al., 2017; Dai et al., 2019b). Optimizing such gate selectivity requires sequence-level objectives, including factual alignment and coherence, that are not directly differentiable (Ranzato et al., 2016; Paulus et al., 2018). Building on prior reinforcement learning and preference optimization approaches (Ouyang et al., 2022; Rafailov et al., 2023a), we formulate memory updates as a policy learning problem and optimize them using GRPO (Shao et al., 2024). During inference, the learned gating policy remains fixed, while memory states continue to evolve through adaptive updates at each decoding step. This formulation treats faithfulness hallucination from the lens of memory regulation failure during generation rather than solely a failure of source compression. Our contributions are as follows:

- We introduce REMEMORIZE, a framework for faithful clinical summarization that models factual inconsistency as a failure of memory regulation during autoregressive generation and addresses it through reward-guided training of a recurrent memory that learns what information to retain and how to ground generation in source evidence.
- We propose a learnable GRU-style recurrent memory with a memory-conditioned scalar gate that operates as a lightweight external module over a frozen transformer and evolves across both source encoding and decoding.
- We design a composite reward function combining entailment-based factual alignment, coherence, and hallucination penalties, allowing gate selectivity to be optimized via GRPO without token-level annotations or retrieval augmentation.
- We evaluate on the MIMIC-IV-Ext-BHC and SOAP benchmarks, demonstrating consistent improvements in factuality over strong domain-specific and general-purpose baselines, with ablations confirming the importance of both adaptive memory control and reward-driven optimization.

2 Related Work

Clinical Text Summarization. Recent work in clinical summarization uses LLMs to generate structured summaries from patient data. Approaches for improving factual reliability include verifier-based alignment, claim-level supervision, and structured generation pipelines (Liu et al., 2026; Piya and Beheshti, 2026; Yuan et al., 2025). Other methods address long and noisy inputs through context selection, routing under constrained token budgets (Maynez et al., 2020), and temporal or graph-based grounding for longitudinal clinical data (Li et al., 2025a). Despite these advances, maintaining factual consistency in long clinical narratives remains difficult. Most existing approaches rely on external mechanisms such as retrieval, verification, or post-hoc correction, while generation itself continues to depend on fixed transformer attention dynamics. Consequently, they provide limited control over how information is retained during autoregressive decoding, contributing to inefficient context utilization and factual drift.

Factual Consistency under Long-Context Generation. Factual consistency between generated summaries and source clinical records remains a major challenge in long-context generation, where unsupported statements can affect downstream clinical decision-making. LLMs may generate fluent yet ungrounded content even when relevant evidence is available in the input (Ji et al., 2023; Maynez et al., 2020; McKenna et al., 2023). Prior work suggests that these failures stem not only from missing information but also from limitations in how models retain and utilize evidence during autoregressive generation, particularly when relevant content is distributed across long and noisy documents (Liu et al., 2024). Existing approaches mitigate hallucinations through post-hoc verification, uncertainty estimation, multi-step reasoning, and context compression strategies (Farquhar et al., 2024; Ke et al., 2024; Liu et al., 2025; Frantar and Alistarh, 2023; Dettmers et al., 2022; Goyal et al., 2020; Hinton et al., 2015). However, these methods do not provide explicit control over how evidence is retained during decoding. In contrast, REMEMORIZE introduces a learnable memory control mechanism that dynamically regulates the preservation and updating of source evidence throughout generation.

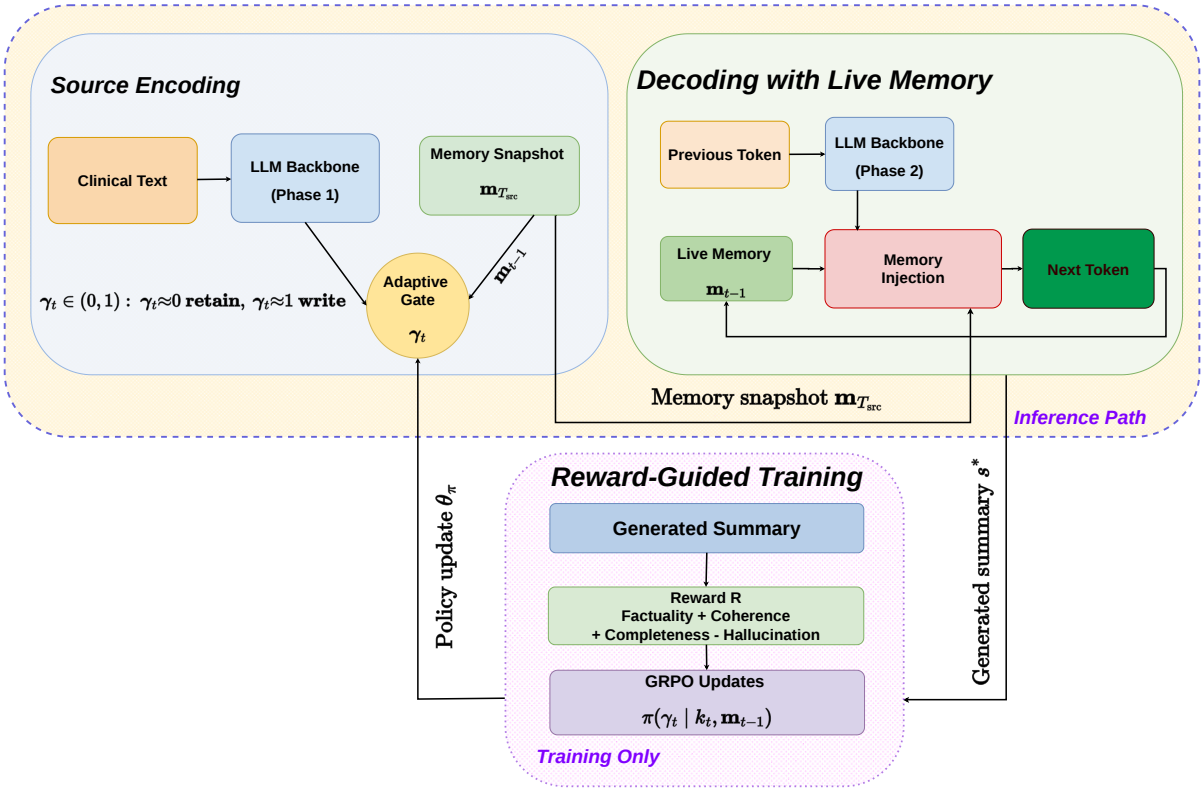

Figure 1: Overview of the REMEMORIZE framework. During training, the model generates provisional summaries evaluated using a composite reward function. During inference, the policy is frozen.

Memory Mechanisms in Sequence Models. Memory-augmented sequence models have been proposed to extend context and capture long-range dependencies. Architectures such as Transformer-XL (Dai et al., 2019a) and Compressive Transformer (Rae et al., 2019) incorporate recurrent memory, while models including Memformer (Wu et al., 2022) and MART (Lei et al., 2020) employ gated memory updates within transformer layers. In summarization, approaches such as Sliding Selector Networks (Cui and Hu, 2021), NeuSum (Zhou et al., 2018), and MemSum (Gu et al., 2022) use memory to track extraction history and reduce redundancy, while recent work on long-dialogue summarization demonstrates that iterative memory-aware strategies improve coherence and factual consistency (Wang et al., 2024). However, most existing methods rely on fixed or heuristic memory updates and provide limited control over how information is retained during generation. In systems such as MemSum (Gu et al., 2022) and Sliding Selector Networks (Cui and Hu, 2021), memory largely functions as a static or externally maintained representation during decoding, limiting adaptability under long and noisy contexts. In contrast, REMEMORIZE maintains a continuously updated memory state through a learnable gating policy during generation.

Reinforcement Learning for Summarization. RL has become a prominent framework for aligning language models with high-level objectives such as relevance, coherence, and factual consistency. Methods including Direct Preference Optimization (Rafailov et al., 2023b), RRHF (Yuan et al., 2023), and RLAIF (Lee et al., 2024) demonstrate that preference-based supervision can guide generation without explicit reward models, while self-rewarding approaches support iterative refinement (Yuan et al., 2024). In summarization, these methods enable optimization of properties that are difficult to capture through token-level likelihood objectives alone (Ryu et al., 2024; Li et al., 2025b; Song et al., 2025; Yadav et al., 2025; Yang et al., 2025). Related work also explores agentic frameworks that decompose generation into drafting, critique, and revision stages (Sudarshan et al., 2024; Tu et al., 2025; Piya and Beheshti, 2025b). However, these approaches primarily optimize outputs before or after decoding rather than regulating how information is retained during generation. In contrast, REMEMORIZE learns a memory-update

policy that directly controls information retention within the decoding process.

3 Methodology

REMEMORIZE addresses a core limitation of autoregressive generation over long clinical text: evidence distributed across thousands of tokens is progressively lost through attention dilution, leading to factually inconsistent summaries. To address this, the framework maintains a latent memory state $\mathbf{m}_t$ that is initialized during source encoding and updated throughout decoding under a learned gating policy. As shown in Figure 1, the framework consists of three components: *Source Encoding*, *Decoding with Live Memory*, and *Reward-Guided Training*. The full procedure is summarized in Algorithm 1 (Appendix A).

Problem Formulation. Before discussing the three components, consider an input document d. The model generates a summary s^* while maintaining a latent memory state $\mathbf{m}_t$ across source encoding and autoregressive decoding. At each step t, the LLM hidden state $h_t \in \mathbb{R}^{d_h}$ is projected into key and value representations $k_t \in \mathbb{R}^{d_k}$ and $v_t \in \mathbb{R}^{d_v}$, while a gating policy π_{θ_π} produces coefficients $\gamma_t \in (0, 1)$ controlling memory updates. Source-encoding steps are indexed by $t \in \{1, \ldots, T_{\text{src}}\}$, while decoding steps are indexed by t'.

3.1 Source Encoding

Standard attention struggles to integrate evidence distributed across long clinical notes (Liu et al., 2024; Khandelwal et al., 2018). Rather than relying on retrieval augmentation (Borgeaud et al., 2022; Ram et al., 2023) or context extension (Dai et al., 2019a), REMEMORIZE compresses source evidence into a persistent latent state before generation.

Memory State and Update Rule. The memory state $\mathbf{m}_t \in \mathbb{R}^{d_m}$ evolves through a GRU-style gated update (Cho et al., 2014; Chung et al., 2014), chosen for selective retention under a compact parameter budget:

$$\mathbf{m}_t = (1-\gamma_t)\, \mathbf{m}_{t-1} + \gamma_t \, \tanh(\text{write_proj}(v_t)) \,, \tag{1}$$

where $v_t \in \mathbb{R}^{d_v}$ is the value representation at step t, and $\text{write_proj} : \mathbb{R}^{d_v} \to \mathbb{R}^{d_m}$ maps token values into the memory space. The scalar gate $\gamma_t \in (0, 1)$ controls memory retention and overwrite. The memory is initialized as $\mathbf{m}_0 = \mathbf{0}$. The computation of γ_t is described below.

Memory-Conditioned Gating. Many memory-augmented architectures condition write decisions primarily on the incoming token, with limited dependence on the existing memory state (Wu et al., 2022; Lei et al., 2020). This can produce redundant updates and gradual dilution of salient evidence in long clinical documents.

To address this, γ_t is computed using a two-layer MLP conditioned jointly on the current token key $k_t \in \mathbb{R}^{d_k}$ and a linear projection of the previous memory state $\mathbf{m}_{t-1}$:

$$\gamma_t = \sigma(\text{MLP}([k_t \,;\, \text{proj}(\mathbf{m}_{t-1})])) \,, \tag{2}$$

where $\text{proj} : \mathbb{R}^{d_m} \to \mathbb{R}^{d_k}$ maps memory representations to the key dimension prior to concatenation, and the MLP produces a scalar logit whose sigmoid defines γ_t. Conditioning on both k_t and $\mathbf{m}_{t-1}$ allows the model to suppress updates for information already represented in memory and focus on novel content. The MLP output bias is initialized to -2.0, giving $\sigma(-2) \approx 0.12$, to keep the gate near-closed at the start of training and reduce early saturation.

In **Phase 1**, the model processes the source document token-by-token in a causal forward pass (Rae et al., 2019; Wu et al., 2022). At each source token $t \in \{1, \ldots, T_{\text{src}}\}$, γ_t is computed using Eq. 2 and $\mathbf{m}_t$ is updated using Eq. 1. After the final source token, $\mathbf{m}_{T_{\text{src}}}$ provides a fixed-size representation of the encoded source context.

3.2 Decoding with Live Memory

In **Phase 2**, during generation, the memory state $\mathbf{m}_{t'-1}$ is injected into the decoder hidden state $h_{t'} \in \mathbb{R}^{d_h}$ through a learned gated projection:

$$\tilde{h}_{t'} = h_{t'} + \sigma(\text{mem_gate}(h_{t'})) \cdot \text{mem_proj}(\mathbf{m}_{t'-1}), \tag{3}$$

where $\text{mem_proj} : \mathbb{R}^{d_m} \to \mathbb{R}^{d_h}$ projects memory into the hidden-state space, and $\text{mem_gate} : \mathbb{R}^{d_h} \to \mathbb{R}$ produces a scalar gate. The memory state continues to evolve during decoding. After token generation from $\tilde{h}_{t'}$, the pre-injection hidden state $h_{t'}$ produces $k_{t'}$ and $v_{t'}$, which drive the GRU update in Eqs. 1 and 2, advancing the memory to $\mathbf{m}_{t'}$ before the next decoding step.

3.3 Reward-Guided Training

To control the memory update schedule γ_t during source encoding and decoding, a gating policy π_{θ_π}

is trained, which is the only trainable component in the RL training phase; the LLM backbone and write projection remain frozen. Rewards are therefore defined over complete generation trajectories relative to the source document.

Factual summarization requires reference-free rewards that remain sensitive to hallucination. Prior work has shown that NLI entailment probability is a reliable proxy for factual consistency (Maynez et al., 2020; Honovich et al., 2022), while contradiction probability captures source-conflicting content (Honovich et al., 2022). Following these, we define a composite reward:

$$R = \alpha\, R_{\text{factual}} + \beta\, R_{\text{coherence}} + \lambda\, R_{\text{completeness}} - \delta\, R_{\text{hallucination}}, \tag{4}$$

where $\alpha, \beta, \lambda, \delta > 0$ are scalar weights (values in Section 4).

Factuality and Hallucination. $R_{\text{factual}} = P_{\text{NLI}}(\textit{entailment} \mid d, s^*)$ and $R_{\text{hallucination}} = P_{\text{NLI}}(\textit{contradiction} \mid d, s^*)$ are computed during training using a DeBERTa-v3-Large cross-encoder (He et al., 2021). Because NLI predictions are not complementary, both terms are required: generated content may be source-neutral without being contradicted. We therefore use contradiction probability as a proxy for hallucination (Honovich et al., 2022). Factuality evaluation uses MiniCheck-Flan-T5-Large (Tang et al., 2024).

Coherence. Let $\{s'_1, \ldots, s'_n\}$ denote the sentences of s^* and $\mathbf{e}'_i$ their embeddings. $R_{\text{coherence}}$ measures local discourse consistency as the mean cosine similarity between adjacent sentence embeddings, following prior work on reward-based summarization optimization (Ryu et al., 2024):

$$R_{\text{coherence}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \cos(\mathbf{e}'_i,\, \mathbf{e}'_{i+1}). \tag{5}$$

Completeness. Let $\{s_1, \ldots, s_m\}$ denote the source-document sentences and $\mathbf{e}_j$ their embeddings, and let n denote the number of generated sentences. $R_{\text{completeness}}$ measures semantic source coverage by matching each source sentence to its closest generated sentence (Zhang et al., 2020):

$$R_{\text{completeness}} = \frac{1}{m} \sum_{j=1}^{m} \max_{i \in \{1,\ldots,n\}} \cos(\mathbf{e}_j,\, \mathbf{e}'_i). \tag{6}$$

Adaptive Reward Normalization. Because reward scales vary across documents (Ouyang et al., 2022), all reward components are normalized using an EMA z-score with decay η_{ema}=0.1, clipped to $[-2.5,\, 2.5]$, stabilizing policy-gradient updates across varying reward magnitudes.

Policy Optimization. Because token-level gate decisions are not differentiable with respect to the sequence-level reward R (Ranzato et al., 2016; Paulus et al., 2018), we optimize the policy using GRPO (Shao et al., 2024), which estimates advantages through within-group reward normalization. For each document d, the policy samples G candidate summaries and computes:

$$\hat{A}_g = \frac{R_g - \mu_G}{\sigma_G + \epsilon_0}, \tag{7}$$

where R_g is the composite reward for the g-th sampled summary, $\mu_G = \frac{1}{G}\sum_{g=1}^{G} R_g$, $\sigma_G = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(R_g - \mu_G)^2}$, and ϵ_0 is a numerical stabilizer.

Policy updates follow the clipped surrogate objective (Schulman et al., 2017):

$$\mathcal{L}_{\text{RL}} = \mathbb{E}\Big[\min\Big(\rho_g \hat{A}_g,\; \text{clip}(\rho_g, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\, \hat{A}_g\Big)\Big], \tag{8}$$

where $\rho_g = \pi_{\theta_\pi}(s^*_g \mid d) \,/\, \pi^{\text{old}}_{\theta_\pi}(s^*_g \mid d)$ is the importance ratio and ϵ_{clip} controls the trust-region size. Training halts early within each epoch when the mean approximate KL divergence between π_{θ_π} and $\pi^{\text{old}}_{\theta_\pi}$ exceeds δ_{KL}, reducing degenerate reward-maximizing behavior during RL fine-tuning (Ouyang et al., 2022). The primary departure from standard GRPO is the multi-component reward (Eq. 4) with per-component EMA normalization.

3.4 Training Objective

Training alternates between optimizing memory and policy parameters (Appendix A, Algorithm 1). The memory interface parameters θ_M are updated using AdamW (Loshchilov and Hutter, 2017) with:

$$\mathcal{L}_{\text{mem}} = \mathcal{L}_{\text{sum}} + \lambda_{\text{mem}}\, \mathcal{L}_{\text{aux}}, \tag{9}$$

where $\mathcal{L}_{\text{sum}}$ is the cross-entropy summarization loss and λ_{mem} is a regularization coefficient selected empirically (Appendix F). The auxiliary reconstruction loss provides token-level supervision for the

memory representation independently of the policy gradient signal:

$$\mathcal{L}_{\text{aux}} = \frac{1}{T_{\text{src}}} \sum_{t=1}^{T_{\text{src}}} \left\| \tanh(\text{write_proj}(v_t)) - \mathbf{m}_{T_{\text{src}}} \right\|^2, \tag{10}$$

encouraging token projections to remain compatible with the final encoded memory state $\mathbf{m}_{T_{\text{src}}}$, and separating memory representation learning from the GRPO-optimized gating policy γ_t. The policy parameters θ_π are optimized separately via GRPO to maximize the composite reward R.

4 Experimental Setup

Datasets We evaluate on two clinical summarization benchmarks. **MIMIC-IV-Ext-BHC** (Aali et al., 2024) contains 279,033 clinical notes paired with brief hospital course summaries (avg. 2,267 tokens); we use 1,000/500/500 for train/val/test. **SOAP** (Neupane et al., 2024) comprises 1,473 patient–doctor conversations from MTS-Dialog (Ben Abacha et al., 2023) annotated with structured SOAP summaries; we use 1,023/150/300 for train/val/test.

Evaluation Metrics Generated summaries are evaluated across five dimensions: lexical overlap, semantic similarity, factual faithfulness, and LLM-based judgment and human-in-the-loop experiment. **Lexical and Semantic.** BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore F1 (Zhang et al., 2020). **Faithfulness.** MiniCheck (Tang et al., 2024) decomposes each summary into atomic claims assessed for source entailment, yielding **Factuality** (Fact.↑) and **Hallucination** (Hall.↓) (Appendix I); the annotation prompt used for hallucination span identification is provided in Appendix B. **LLM-as-a-Judge.** An ensemble of LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen2.5-7B-Instruct scores each summary on Hallucination (↓), Factual Consistency (↑), Completeness (↑), and Coherence (↑); scores are averaged across judges (Appendix C). **Human Evaluation.** A blinded pairwise study (n=20) compared REMEMORIZE against the strongest baseline across three clinical cases, with participants rating hallucination severity (Appendix E). The survey remains publicly accessible.[1]

[1] https://docs.google.com/forms/d/e/1FAIpQLSeHB27VYkQrA7scQFFkYGLsj3Nt7XWHmoKxsZ-cW_9xkuonXA/viewform?usp=header

Baselines We evaluate REMEMORIZE against a diverse set of instruction-tuned LLMs spanning parameter scales, architectures, and levels of domain specialization, enabling controlled assessment of whether improvements in factual consistency arise from adaptive memory control rather than model capacity or pre-training.

General-purpose models. We include LLaMA-3.2 models at 1B and 3B scales (Touvron et al., 2023) to assess scaling within a single architecture, along with Qwen2.5 models at 3B and 7B (Bai et al., 2023), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Phi-3.5-mini-Instruct (Abdin et al., 2024), and Gemma-2-2B-IT (Team et al., 2024), covering competitive instruction-tuned models in the 2B–7B range. **Domain-adapted models.** We include BioMistral-7B (Labrak et al., 2024), MedAlpaca-7B (Han et al., 2023), and SumOrchestra-8B (Mei et al., 2026) to evaluate the impact of biomedical pre-training, clinical instruction-tuning, and domain-specific reward learning, respectively. **Additional baselines.** We further include DeepSeek-R1-8B (DeepSeek-AI, 2024) and MemSum-PubMed (Gu et al., 2022) to assess the roles of reasoning and explicit memory mechanisms without adaptive control. Full descriptions of all baseline models are provided in Appendix D.

Implementation Details REMEMORIZE uses a frozen `LLaMA-3.1-8B-Instruct` backbone with a GRU-based memory module trained via GRPO. We use a rollout group size of G=4 for stable advantage estimation on long clinical documents. Full hyperparameters and reward-weight sensitivity analyses are provided in Appendices F and J.

5 Results

Table 1 reports automatic evaluation results. Across both datasets, REMEMORIZE achieves the highest factuality scores while remaining competitive on hallucination and summary quality metrics. On MIMIC-IV-Ext-BHC, REMEMORIZE achieves the strongest factuality score at 74.3, outperforming DeepSeek-R1-8B (70.8) by 3.5 points. It also attains the best R-1, R-2, R-L, and BERTScore F1 values, improving the strongest baseline BERTScore from 82.7 to 83.9. On SOAP, REMEMORIZE again achieves the highest factuality score at 71.8, exceeding the strongest baseline, LLaMA-3.2-1B (65.2), by 6.6 points. It further achieves the best R-1 and R-2 scores, while main-

Table 1: Summary generation performance on MIMIC-IV-Ext-BHC (n=500) and SOAP (n=300). R-1/R-2/R-L: ROUGE-1/2/L (lexical overlap); BS-F1: BERTScore-F1 (semantic similarity); Fact.: MiniCheck factuality score; Hall.: NLI-based hallucination rate. ↑ higher is better, ↓ lower is better. Values: mean$_{\pm \text{std}}$ (1000 bootstrap iterations). Bold: best per column.

	MIMIC-IV-Ext-BHC						SOAP					
Model	**R-1↑**	**R-2↑**	**R-L↑**	**BS-F1↑**	**Fact.↑**	**Hall.↓**	**R-1↑**	**R-2↑**	**R-L↑**	**BS-F1↑**	**Fact.↑**	**Hall.↓**
LLaMA-3.2-1B	23.4	4.2	13.1	80.9	66.6	14.7	22.2	5.8	13.2	82.6	65.2	14.7
LLaMA-3.2-3B	21.2	3.9	11.7	51.8	52.7	14.1	30.3	14.7	27.7	86.0	46.2	8.5
Qwen2.5-3B	28.3	6.1	15.9	82.7	62.3	12.8	28.0	16.9	**29.7**	87.3	50.3	6.7
Mistral-7B	28.6	6.1	15.8	55.4	51.2	7.9	27.5	14.1	25.7	83.6	61.4	6.5
Qwen2.5-7B	29.3	7.0	17.2	82.1	58.4	**7.7**	26.0	16.5	29.0	83.8	51.9	6.0
BioMistral-7B	28.1	7.0	17.0	82.6	69.3	15.8	26.1	8.1	16.5	82.6	50.9	6.9
MedAlpaca-7B	22.3	4.5	13.0	82.4	53.8	15.1	26.2	10.3	18.8	83.0	47.2	6.7
Phi-3.5-mini	26.5	5.2	14.9	82.7	53.5	8.3	29.9	16.6	28.2	**87.5**	60.4	6.8
Gemma-2-2B	28.3	6.6	16.0	82.3	65.1	23.4	28.6	9.6	19.2	83.9	60.9	7.8
DeepSeek-R1-8B	13.5	1.2	8.7	80.6	70.8	28.3	10.8	1.1	7.1	80.2	61.3	8.5
MemSum-PubMed	19.0	5.2	11.9	81.0	58.3	33.3	18.2	7.8	13.9	82.5	52.2	24.8
ReMemorize	$\mathbf{31.1}_{\pm 0.3}$	$\mathbf{7.1}_{\pm 0.2}$	$\mathbf{18.8}_{\pm 0.2}$	$\mathbf{83.9}_{\pm 0.1}$	$\mathbf{74.3}_{\pm 1.7}$	$8.2_{\pm 1.1}$	$\mathbf{30.7}_{\pm 0.5}$	$\mathbf{18.2}_{\pm 0.3}$	$17.9_{\pm 0.3}$	$84.2_{\pm 0.1}$	$\mathbf{71.8}_{\pm 2.3}$	$\mathbf{5.9}_{\pm 1.2}$

Table 2: LLM-as-a-Judge evaluation (ensemble of LLaMA-3.2-3B, Mistral-7B, and Qwen2.5-7B). All scores on a 1–5 scale.

	MIMIC-IV-Ext-BHC				SOAP			
Model	**Hallucination↓**	**Factual Consist.↑**	**Completeness↑**	**Coherence↑**	**Hallucination↓**	**Factual Consist.↑**	**Completeness↑**	**Coherence↑**
LLaMA-3.2-1B	$1.10_{\pm 0.33}$	$4.29_{\pm 0.35}$	$3.70_{\pm 0.42}$	$4.38_{\pm 0.39}$	$1.27_{\pm 0.54}$	$2.71_{\pm 1.53}$	$2.38_{\pm 1.31}$	$3.48_{\pm 1.36}$
LLaMA-3.2-3B	$1.34_{\pm 0.57}$	$4.49_{\pm 0.30}$	$3.91_{\pm 0.46}$	$4.66_{\pm 0.35}$	$1.19_{\pm 0.40}$	$4.55_{\pm 0.31}$	$3.97_{\pm 0.68}$	$4.77_{\pm 0.32}$
Qwen2.5-3B	$1.14_{\pm 0.39}$	$4.60_{\pm 0.23}$	$4.21_{\pm 0.39}$	$4.82_{\pm 0.25}$	$1.06_{\pm 0.21}$	$4.59_{\pm 0.41}$	$4.12_{\pm 0.57}$	$4.81_{\pm 0.35}$
Qwen2.5-7B	$1.21_{\pm 0.46}$	$4.66_{\pm 0.26}$	$4.14_{\pm 0.37}$	$4.40_{\pm 0.21}$	$1.35_{\pm 0.52}$	$4.55_{\pm 0.33}$	$4.28_{\pm 0.43}$	$4.67_{\pm 0.24}$
Mistral-7B	$1.45_{\pm 0.53}$	$4.48_{\pm 0.22}$	$4.02_{\pm 0.36}$	$4.81_{\pm 0.26}$	$1.65_{\pm 0.25}$	$4.55_{\pm 0.26}$	$4.09_{\pm 0.44}$	$4.83_{\pm 0.26}$
BioMistral-7B	$1.26_{\pm 0.48}$	$4.23_{\pm 0.63}$	$3.58_{\pm 0.63}$	$4.55_{\pm 0.63}$	$1.48_{\pm 0.54}$	$4.24_{\pm 0.67}$	$3.79_{\pm 0.72}$	$4.63_{\pm 0.57}$
MedAlpaca-7B	$1.32_{\pm 0.60}$	$4.26_{\pm 0.71}$	$3.44_{\pm 0.83}$	$4.52_{\pm 0.67}$	$1.48_{\pm 0.63}$	$4.34_{\pm 0.75}$	$3.85_{\pm 0.92}$	$4.56_{\pm 0.63}$
Phi-3.5-mini	$1.35_{\pm 0.52}$	$4.51_{\pm 0.24}$	$4.06_{\pm 0.37}$	$4.81_{\pm 0.25}$	$1.36_{\pm 0.48}$	$4.50_{\pm 0.30}$	$3.79_{\pm 0.52}$	$4.78_{\pm 0.27}$
Gemma-2-2B	$1.49_{\pm 0.42}$	$4.31_{\pm 0.45}$	$3.46_{\pm 0.54}$	$4.65_{\pm 0.52}$	$1.07_{\pm 0.27}$	$4.55_{\pm 0.54}$	$3.80_{\pm 1.01}$	$4.71_{\pm 0.54}$
Flan-T5-XXL	$1.71_{\pm 0.81}$	$4.57_{\pm 0.49}$	$3.52_{\pm 0.84}$	$4.78_{\pm 0.38}$	$1.71_{\pm 0.70}$	$3.50_{\pm 1.21}$	$2.58_{\pm 0.97}$	$4.11_{\pm 1.08}$
SummOrchestra-8B	$1.36_{\pm 0.58}$	$4.41_{\pm 0.48}$	$3.33_{\pm 0.63}$	$4.76_{\pm 0.43}$	$1.60_{\pm 0.62}$	$4.47_{\pm 0.46}$	$3.82_{\pm 0.69}$	$4.81_{\pm 0.32}$
ReMemorize	$\mathbf{1.06}_{\pm 0.53}$	$\mathbf{4.68}_{\pm 0.27}$	$\mathbf{4.29}_{\pm 0.44}$	$\mathbf{4.83}_{\pm 0.44}$	$\mathbf{1.30}_{\pm 0.52}$	$\mathbf{4.68}_{\pm 0.37}$	$\mathbf{4.32}_{\pm 0.53}$	$\mathbf{4.86}_{\pm 0.34}$

taining a hallucination rate of 5.9, comparable to the strongest baseline result of 6.0 from Qwen2.5-7B. However, it does not achieve the strongest R-L or BERTScore F1 values on SOAP. MemSum-PubMed, an extractive baseline, achieves factuality scores of 58.3 and 52.2 on MIMIC-IV-Ext-BHC and SOAP, respectively. Despite generating summaries entirely from source sentences, its factuality remains below stronger generative baselines, consistent with prior findings that extractive selection alone does not guarantee claim-level faithfulness under noisy or de-identified clinical text (Maynez et al., 2020). Overall, these results suggest that the primary advantage of REMEMORIZE lies in improving factual grounding, with broader gains in lexical and semantic quality that are strongest on MIMIC-IV-Ext-BHC and more selective on SOAP. A qualitative hallucination case study illustrating the types of errors suppressed by the memory module is provided in Appendix L.

Table 2 presents LLM-as-a-judge evaluation results. REMEMORIZE achieves the strongest overall performance, particularly in factual consistency and completeness. On MIMIC-IV-Ext-BHC, it attains the best scores across all four dimensions, including factual consistency (4.68 vs. 4.66), hallucination (1.06 vs. 1.10), completeness (4.29 vs. 4.21), and coherence (4.83 vs. 4.82). On SOAP, it again achieves the highest factual consistency (4.68 vs. 4.59), completeness (4.32 vs. 4.28), and coherence (4.86 vs. 4.83). Its hallucination score is slightly higher than the strongest baseline (1.30 vs. 1.06), indicating a modest trade-off between hallucination suppression and overall summary quality. SumOrchestra-8B, which applies reward-based optimization for clinical dialogue summarization, achieves factuality scores of 4.41 on MIMIC-IV-Ext-BHC and 4.47 on SOAP, remaining below REMEMORIZE on both datasets. This suggests that reward optimization alone may be insufficient

Table 3: **Human evaluation** (n=20 annotators) of hallucination severity across clinical domains. Correct Guess Accuracy (↑): proportion identifying REMEMORIZE as the less hallucinated summary (exact binomial test vs. chance). Hallucination severity rated on a 1–5 ordinal scale (↓). All Wilcoxon signed-rank tests: $p \leq 0.001$; r_{rb}: rank-biserial effect size.

Clinical Domain	**Correct Guess Accuracy (%)** ↑	**Mean Hall. Severity** ↓ *Vanilla LLM* / REMEMORIZE	**Effect Size** (r_{rb})	**Dominance** p
Gastroenterology	90.0	3.95 / **1.95**	0.981	<0.001
Cardiology	90.0	3.50 / **1.80**	0.819	<0.001
General Surgery	100.0	3.50 / **1.55**	0.847	<0.001

Table 4: Backbone generalizability of REMEMORIZE. Each backbone is evaluated in two configurations: without the memory module (backbone only) and with it (REMEMORIZE). Higher is better for all.

Backbone	**Configuration**	**Dataset**	**R-1**↑	**R-2**↑	**R-L**↑	**BS-F1**↑	**Fact.**↑
Mistral-7B	Backbone only	MIMIC	28.6	6.1	15.8	55.4	51.2
	REMEMORIZE	MIMIC	**29.1**	**6.8**	**16.3**	**73.9**	**64.3**
	Backbone only	SOAP	27.5	14.1	**25.7**	83.6	61.4
	REMEMORIZE	SOAP	**28.7**	**16.2**	19.9	**83.8**	**69.8**
LLaMA-3.1-8B	Backbone only	MIMIC	**31.3**	**7.5**	17.8	83.2	58.8
	REMEMORIZE	MIMIC	31.1	7.1	**18.8**	**83.9**	**74.3**
	Backbone only	SOAP	**44.3**	**18.6**	**31.6**	**87.9**	55.3
	REMEMORIZE	SOAP	30.7	18.2	17.9	84.2	**71.8**

without explicit adaptive memory control in long-document summarization settings.

Table 3 reports human evaluation across three clinical domains. Annotators consistently identified REMEMORIZE summaries as less hallucinated than vanilla LLM outputs, with agreement rates of 90% in gastroenterology and cardiology, and 100% in general surgery. Hallucination severity ratings were significantly lower across all domains ($p \leq 0.001$), with large rank-biserial correlations indicating strong effect sizes. Additional participant and experimental details are provided in Appendix E.

Backbone Generalizability & Ablation To evaluate generalizability beyond a single pretrained model family, we introduce Mistral-7B alongside LLaMA-3.1-8B while keeping the memory architecture and training procedure fixed. Across both datasets, REMEMORIZE consistently improves factuality over the corresponding backbone-only baselines: +13.1/+8.4 for Mistral-7B and +15.5/+16.5 for LLaMA-3.1-8B on MIMIC-IV-Ext-BHC and SOAP, respectively, suggesting that the recurrent memory mechanism transfers across distinct decoder-only LLM families. We further perform targeted ablations on memory injection, adaptive gating, Phase-2 decoding updates, and reinforcement learning. Full settings and analyses are provided in Table 4 and Appendices G, H, and K.

6 Conclusion

We presented REMEMORIZE, a framework that introduces explicit control over memory updates during autoregressive generation through gated recurrent memory and reinforcement learning. Across two clinical summarization benchmarks, the method consistently improves factual consistency while maintaining competitive lexical and semantic quality, achieving gains of up to +6.6 factuality points over baselines. Ablation studies further demonstrate that both adaptive memory updates and reward-driven policy optimization are necessary for effective memory regulation. More broadly, our results suggest that improving long-context generation may require moving beyond fixed attention dynamics toward explicit mechanisms for controlling information retention during decoding. These findings suggest that explicit memory regulation may provide a scalable approach for improving factual grounding in clinical summarization tasks.

7 Limitations

Our evaluation is limited to two clinical summarization benchmarks, and the generalizability of the proposed memory control mechanism to other domains and tasks remains unexplored. That said, the recurrent memory architecture and GRPO training objective are not inherently domain-specific; the

module imposes no clinical priors and can in principle be applied to other long-context reasoning. The reward function relies on NLI-based scoring, which may introduce bias in specialized clinical settings. We also do not evaluate robustness under real-world deployment conditions. As a potential risk, remaining hallucinated clinical summaries, if acted upon without human oversight, could contribute to clinical decision errors. Our framework, however, is intended for research purposes and could be deployed in clinical settings only after further validation.

References

Asad Aali, Peter H. Van Ness, Michael Simonov, and Benjamin A. Goldstein. 2024. MIMIC-IV-Ext-BHC: A benchmark dataset for clinical text summarization. *arXiv preprint arXiv:2405.05502*.

Marah Abdin, Sam Jacobs, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sina Asgari and 1 others. 2025. A framework for evaluating and mitigating hallucinations in large language models. *arXiv preprint*.

Jinze Bai, Shuai Bai, Yunfei Yang, and et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Asma Ben Abacha, Wen-wai Yim, George Adams, Neal Donahue, and Hong Yu. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Yung-Sung Chuang, Linlu Xie, Cheng-Yu Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Peng Cui and Le Hu. 2021. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5881–5891.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019a. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019b. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2978–2988.

DeepSeek-AI. 2024. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2401.XXXX*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one shot. In *Proceedings of the 40th International Conference on Machine Learning*.

Saurabh Goyal, Aditya Romesh Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination. In *Proceedings of the 37th International Conference on Machine Learning*.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nianlong Gu, Elliott Ash, and Richard H. R. Hahnloser. 2022. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6507–6522.

Tianyi Han and 1 others. 2023. Medalpaca: An open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Ziwei Ji and 1 others. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, and et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ping Ke and 1 others. 2024. Enhancing clinical reasoning with multi-agent frameworks. In *arXiv preprint*.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294.

Younes Labrak and 1 others. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2024. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML 2024.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. 2020. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614.

Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. 2025a. Clicare: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. *Preprint*, arXiv:2507.22533.

Yang Li and 1 others. 2025b. Topic-aware GRPO for multi-document summarization. *arXiv preprint*.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 641–647.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Shenling Liu, Yang Gao, ShaSha Li, PanCheng Wang, and Ting Wang. 2025. A hallucination detection and mitigation framework for faithful text summarization using llms. *Scientific Reports*.

Weixin Liu, Congning Ni, Qingyuan Song, Susannah L. Rose, Christopher Symons, Murat Kantarcioglu, Bradley A. Malin, and Zhijun Yin. 2026. Veridpo: Evidence-aware alignment for clinical summarization via claim verification and direct preference optimization. *Preprint*, arXiv:2603.10494.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback, 2023. *URL https://arxiv.org/abs/2303.17651*, 2303.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Xiaoyong Mei, Tingting Zuo, Da Chen, Guangyu Hu, Xiangyu Wen, Chao Duan, Mingyan Zhang, and Fudan Zheng. 2026. Beyond overlap metrics: Rewarding reasoning and preferences for faithful

multi-role dialogue summarization. *arXiv preprint arXiv:2604.17188*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Amir Moeini, Jiuqi Wang, Jacob Beck, Ethan Blaser, Shimon Whiteson, Rohan Chandra, and Shangtong Zhang. 2025. A survey of in-context reinforcement learning. *arXiv preprint arXiv:2502.07978*.

Aishik Nagar, Yutong Liu, Andy T Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, Guna Pratheep Kalanchiam, Yili Tang, and Robby T Tan. 2025. umedsum: a unified framework for clinical abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2672.

Siddharth Neupane and 1 others. 2024. SOAP medical conversation summaries. Hugging Face Datasets.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: towards llms as operating systems.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph 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 1–22.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Fahmida Liza Piya and Rahmatollah Beheshti. 2025a. Advancing feature extraction in healthcare through the integration of knowledge graphs and large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 29293–29294.

Fahmida Liza Piya and Rahmatollah Beheshti. 2025b. Contextual: Improving clinical text summarization in llms with context-preserving token filtering and knowledge graphs. *arXiv preprint arXiv:2504.16394*.

Fahmida Liza Piya and Rahmatollah Beheshti. 2026. Agenticsum: An agentic inference-time framework for faithful clinical text summarization. *arXiv preprint arXiv:2602.20040*.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023a. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023b. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 2024. Multi-dimensional optimization for text summarization via reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5871, Bangkok, Thailand. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8371–8384.

Jing Song and 1 others. 2025. HVO: Hypervolume optimization for multi-objective text summarization. *arXiv preprint*.

Manjunath Sudarshan and 1 others. 2024. Agentic large language models for multi-disciplinary healthcare summary generation. *arXiv preprint*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822.

Liyan Tang, Philippe Laban, Esin Durmus, Pengfei Liu, Wenya Xu, and Noah A. Smith. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Gemma Team and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu and 1 others. 2025. Towards conversational diagnostic AI: AMIE. *arXiv preprint arXiv:2401.05654*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2024. Recursive summarization enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*.

Qingyang Wu, Zhenzhong Lan, Kun Gu, and Zhou Yu. 2022. Memformer: A memory-augmented transformer for sequence modeling. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: a systematic review with large language models and beyond. *MedRxiv*.

Prashant Yadav and 1 others. 2025. ReLSum: Reinforcement learning for query-relevant product summarization. *arXiv preprint*.

Zheng Yang and 1 others. 2025. Aligning radiology report generation with expert preferences via RLAIF. *arXiv preprint*.

Cheng Yuan, Xinkai Rui, Yongqi Fan, Yawei Fan, Boyang Zhong, Jiacheng Wang, Weiyan Zhang, and Tong Ruan. 2025. Lcds: A logic-controlled discharge summary generation system supporting source attribution and expert review. *Preprint*, arXiv:2507.05319.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Suber, Jason Weston, and 1 others. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback without tears. In *Advances in Neural Information Processing Systems*, volume 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 654–663.

Appendix

A Training Algorithm

Algorithm 1 provides a step-by-step summary of the REMEMORIZE training procedure. The two phases correspond directly to the architecture described in Section 3: Phase 1 encodes the source document into the memory state $\mathbf{m}$ token by token, with the adaptive gate γ_t controlling how much each new key–value pair updates the memory. Phase 2 runs G parallel decoding rollouts, each of which continues updating the memory as new tokens are generated, so the memory state reflects both source context and the partially decoded summary. After all G rollouts are complete, GRPO computes normalized advantages and updates the policy parameters θ_π; the memory parameters θ_M are updated separately via AdamW on the supervised summarization loss plus the auxiliary regularizer.

Algorithm 1: REMEMORIZE Training Procedure

Input: Document d, group size G, KL threshold δ_{KL}
Initialize $\mathbf{m}_0 \leftarrow \mathbf{0}$;
`// Phase 1: Source Encoding`
for $t = 1, \ldots, T_{src}$ **do**
  Compute (k_t, v_t) from LLM hidden states;
  Compute γ_t via Eq. 2;
  Update $\mathbf{m}_t$ via Eq. 1;
`// Phase 2: Decoding with Live Memory`
for $g = 1, \ldots, G$ **do**
  $\mathbf{m}_0^{(g)} \leftarrow \mathbf{m}_{T_{\text{src}}}$;
  for $t' = 1, \ldots, T_{dec}$ **do**
    Inject memory via Eq. 3;
    Sample next token from $\text{lm_head}(\tilde{h}_{t'})$;
    Compute $(k_{t'}, v_{t'})$ from $h_{t'}$;
    Compute $\gamma_{t'}$ via Eq. 2;
    Update $\mathbf{m}_{t'}^{(g)}$ via Eq. 1;
  $R_g \leftarrow R(d, s_g^*)$;
Compute normalized advantages $\hat{A}_g$ via Eq. 7;
Update θ_π via GRPO; stop if $\text{KL} > \delta_{\text{KL}}$;
Update θ_M via AdamW using $\mathcal{L}_{\text{sum}} + 0.1\,\mathcal{L}_{\text{aux}}$;

B Hallucination Annotation Prompt

To identify hallucinated spans in model-generated clinical summaries, we employ a structured evaluation prompt grounded in strict textual entailment. The task requires a language model to verify whether each statement in a generated summary is verifiably supported by the source document. This prompt is adapted from the Lookback Lens (Chuang et al., 2024) framework with additional guardrails for clinical applications.

You will be provided with a clinical source document and a machine-generated summary. Your task is to assess whether the summary is **strictly entailed** by the source document.

Definition of Entailment: A statement is entailed only if it is explicitly and unambiguously stated in the document. Any statement that requires inference, clinical reasoning, or prior medical knowledge should be treated as *Not Entailed*.

Key Instructions:

- DO NOT infer or assume missing information.
- DO NOT use clinical knowledge, reasoning, or common sense.
- DO NOT validate based on plausibility, typicality, or tone.

You must treat the document as the **only source of ground truth**.

Document: {document} **Proposed Summary:** {summary}

Output Format:

- **Entailment Label:** `Entailed` or `Not Entailed`
- **Explanation:** Justify with direct references to the document.
- **Problematic Spans:** List summary phrases not directly supported.

C LLM-as-a-Judge Evaluation Protocol

Judge ensemble. We instantiate three instruction-tuned judge models: LLaMA-3.2-3B-Instruct (Touvron et al., 2023) ($\mathcal{J}_{\text{L}}$), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) ($\mathcal{J}_{\text{M}}$), and Qwen2.5-7B-Instruct (Bai et al., 2023) ($\mathcal{J}_{\text{Q}}$), forming the judge set $\mathcal{J} = \{\mathcal{J}_{\text{L}}, \mathcal{J}_{\text{M}}, \mathcal{J}_{\text{Q}}\}$. Each judge independently evaluates every (source, summary) pair and returns an integer score $s_{j,d}(x) \in \{1, 2, 3, 4, 5\}$ for each evaluation dimension $d \in \mathcal{D} = \{\text{HALL}, \text{FACT}, \text{COMP}, \text{COH}\}$.

Ensemble scoring. To mitigate idiosyncratic biases of any single judge, we aggregate scores via arithmetic mean over the subset of judges $\mathcal{J}^*(x) \subseteq \mathcal{J}$ that produce a parseable response for example x:

$$\bar{s}_d(x) = \frac{1}{|\mathcal{J}^*(x)|} \sum_{j \in \mathcal{J}^*(x)} s_{j,d}(x). \quad (11)$$

Corpus-level scores are obtained by averaging over all N test examples:

$$\bar{S}_d = \frac{1}{N} \sum_{x=1}^{N} \bar{s}_d(x). \quad (12)$$

For interpretability, raw scores are also normalized to the unit interval as $\tilde{s}_{j,d}(x) = (s_{j,d}(x) - 1)/4 \in [0, 1]$.

Parse-failure handling. Judges occasionally produce malformed output. A response is marked as a parse failure if it does not contain all four required score fields in the expected format. In such cases, the corresponding judge is excluded from $\mathcal{J}^*(x)$ for that example only, and the ensemble falls back to the remaining judges. This per-example exclusion ensures that every summary receives a valid ensemble score regardless of individual judge failures.

Decoding configuration. All judges are queried with temperature $\tau = 0.3$ and a maximum output length of 150 tokens to encourage deterministic, concise scoring.

Evaluation prompt. Each judge receives the following structured prompt. The model is explicitly instructed to assess summaries solely with respect to the provided source document, without invoking external medical knowledge, following the source-grounded evaluation protocol of Asgari et al. (2025).

> **Task:** Evaluate the generated medical summary against the source clinical document. Assign an integer score from 1 to 5 for each criterion below.
>
> **Source Document:** {SOURCE_DOCUMENT}
> **Generated Summary:** {GENERATED_SUMMARY}
>
> **Evaluation Criteria (1–5 scale):**
>
> - **Hallucination:** Degree of unsupported or fabricated content (1 = none; 5 = major fabrications)
> - **Factual Consistency:** Faithfulness to the source (1 = highly inaccurate; 5 = fully accurate)
> - **Completeness:** Coverage of core clinical information (1 = key info missing; 5 = fully comprehensive)
> - **Coherence:** Fluency and logical organization (1 = poorly written; 5 = highly coherent)
>
> **Output Format** (no additional text):

```
Hallucination: X
Factual: X
Complete: X
Coherent: X
```

D Baseline Models

We evaluate REMEMORIZE against thirteen baseline systems spanning four categories: general-purpose instruction-tuned language models, biomedical and clinical domain-adapted models, a reasoning-optimised distilled model, and a classical extractive system. All generative baselines are prompted zero-shot with task-appropriate system and user instructions; no fine-tuning is applied to any baseline. Table 5 provides a complete listing.

General-purpose instruction-tuned LLMs. We include eight instruction-tuned general-purpose language models spanning the 1B–11B parameter range in order to cover a diverse set of publicly available architectures, scales, and training strategies at the time of writing. The evaluated models include decoder-only architectures from several widely used model families, including LLaMA-3.2 (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen2.5 (Bai et al., 2023), Phi-3.5-mini (Abdin et al., 2024), and Gemma-2 (Team et al., 2024). These models differ substantially in parameter scale, tokenizer design, instruction-tuning procedures, and long-context behavior, providing a broad evaluation setting for assessing whether improvements in factual consistency arise from the proposed memory mechanism rather than backbone-specific properties. We additionally include Flan-T5-XXL as a strong encoder-decoder sequence-to-sequence baseline. Unlike decoder-only chat models, Flan-T5-XXL performs conditional generation through explicit encoder-decoder attention, offering a complementary architectural comparison for long-document summarization. Including both decoder-only and encoder-decoder models allows evaluation across distinct generation paradigms commonly used in summarization research. All decoder-only models are prompted using standardized system-plus-user chat templates

Table 5: Baseline models evaluated in this work. Scale denotes parameter count; identifiers correspond to Hugging Face model repositories.

Model	Scale	HF Identifier
General-purpose instruction-tuned LLMs		
LLaMA-3.2-1B	1B	`meta-llama/Llama-3.2-1B-Instruct`
LLaMA-3.2-3B	3B	`meta-llama/Llama-3.2-3B-Instruct`
Qwen2.5-3B	3B	`Qwen/Qwen2.5-3B-Instruct`
Mistral-7B	7B	`mistralai/Mistral-7B-Instruct-v0.3`
Qwen2.5-7B	7B	`Qwen/Qwen2.5-7B-Instruct`
Phi-3.5-mini	3.8B	`microsoft/Phi-3.5-mini-instruct`
Gemma-2-2B	2B	`google/gemma-2-2b-it`
Flan-T5-XXL	11B	`google/flan-t5-xxl`
Biomedical and clinical domain-adapted LLMs		
BioMistral-7B	7B	`BioMistral/BioMistral-7B`
MedAlpaca-7B	7B	`medalpaca/medalpaca-7b`
SummOrchestra-8B	8B	`NebulaPixel/SummOrchestra-LLAMA3.1-8B-GRPO-BRL-SAMSUM`
Reasoning-optimised model		
DeepSeek-R1-8B	8B	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`
Extractive baseline		
MemSum-PubMed	N/A	`nianlong/memsum-pubmed-summarization`

consistent with their respective instruction-tuning formats. Flan-T5-XXL receives a short task-specific prefix following standard sequence-to-sequence prompting conventions. No additional fine-tuning, retrieval augmentation, or task-specific adaptation is applied to any baseline model.

Biomedical and clinical domain-adapted LLMs. To evaluate whether domain-specific pre-training and clinical instruction tuning alone are sufficient to improve factual consistency, we include several biomedical and clinically adapted language models designed for medical text generation and reasoning. BioMistral-7B (Labrak et al., 2024) extends Mistral-7B through continual pre-training on a large-scale biomedical corpus derived from PubMed Central, incorporating terminology and distributional patterns specific to biomedical literature. MedAlpaca-7B (Han et al., 2023) is instruction-tuned on medical question-answer data following the Alpaca training framework, emphasizing conversational clinical interaction and medical instruction following. SummOrchestra-8B (Mei et al., 2026) is a clinically oriented model specifically designed for faithful multi-role dialogue summarization using reward-guided optimization strategies. These models provide complementary forms of medical adaptation, including biomedical continual pre-training, clinical instruction tuning, and summarization-oriented reward optimization. Collectively, they enable assessment of whether domain specialization alone is sufficient to match the factual consistency gains achieved by adaptive memory control and reward-guided information retention.

Reasoning-optimised model. We additionally evaluate DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2024), an 8B reasoning-oriented model distilled from DeepSeek-R1 into a LLaMA-3 architecture while retaining chain-of-thought reasoning capabilities. Unlike the clinically adapted baselines, this model emphasizes structured reasoning and multi-step inference rather than biomedical specialization. Including this baseline allows us to assess whether improvements in factual consistency can emerge from stronger reasoning behavior alone, independent of domain adaptation or explicit memory control mechanisms. During evaluation, internal `<think>` reasoning traces are removed from the decoded outputs prior to scoring, and only the final generated summaries are evaluated. This ensures consistent comparison across all baselines and prevents intermediate reasoning tokens from affecting lexical or factuality metrics.

Extractive baseline. MemSum (Gu et al., 2022) is an extractive summarization model trained with reinforcement learning to sequentially select informative sentences. We use the PubMed pre-trained checkpoint (`memsum-pubmed-summarization`), which is the closest available domain match to clinical text. Because MemSum is extractive, its

outputs consist entirely of source sentences and therefore receive high source-recall scores by construction; the factuality scores nonetheless reflect genuine variation in which sentences are selected.

E Human-in-the-Loop Experiment

Three clinical note–summary pairs are drawn from the MIMIC-IV-Ext-BHC test set via stratified selection across clinical service domains: gastroenterology (Case 1), cardiology (Case 2), and general surgery (Case 3). Cases were chosen to represent semantically distinct clinical narratives, ensuring that domain-specific reasoning does not confound cross-case aggregation. All source documents retain the de-identification conventions of MIMIC-IV, with protected health information replaced by bracketed tokens consistent with the original dataset release. Within each case, the two summaries under comparison – one produced by REMEMORIZE and one by Mistral-7B-Instruct – are assigned to labels **A** and **B** via independent randomization, counterbalanced across cases to prevent label-order bias.

Participants. Twenty individuals participated voluntarily in the human evaluation study. All participants held advanced academic degrees, including postgraduate and doctoral-level qualifications (10 MSc/MPH/MEng; 6 MBBS or equivalent; 4 doctoral degrees), yielding a final analytic sample of $n = 20$ respondents.

Procedure. Each respondent evaluated all three cases in a fixed presentation order, with cases separated by page breaks. For each case, the full source clinical note was displayed above two anonymised summaries. Respondents completed the following items per case:

1. **Forced preference** (nominal): Which summary contains more hallucinated content not supported by the source note? (A / B)

2. **Hallucination severity – Summary A** (ordinal, 1–5): Rate the degree of unsupported or fabricated content in Summary A, where 1 = none and 5 = severe.

3. **Hallucination severity – Summary B** (ordinal, 1–5): Rate the degree of unsupported or fabricated content in Summary B.

Statistical analysis. All analyses treat the respondent as the unit of observation. For hallucination severity ratings, the primary test is the Wilcoxon signed-rank test applied to within-respondent paired differences $D_i = s_i^{(\mathrm{R})} - s_i^{(\mathrm{M})}$, where $s_i^{(\mathrm{R})}$ and $s_i^{(\mathrm{M})}$ denote the severity scores assigned to REMEMORIZE and Mistral-7B by respondent i, respectively. We report the test statistic W, two-tailed p-value, and rank-biserial correlation r as a standardized effect-size measure:

$$r \;=\; 1 \;-\; \frac{2W}{n(n+1)/2}, \qquad (13)$$

where W is the smaller of the two rank sums and n denotes the number of respondents with $D_i \neq 0$. The rank-biserial correlation ranges from -1 to $+1$, with positive values indicating that REMEMORIZE receives systematically lower hallucination severity ratings than Mistral-7B. For the forced-preference item, an exact binomial test is applied with null probability $p_0 = 0.5$, where each respondent constitutes one independent Bernoulli trial per case. All tests are two-tailed; the significance threshold is set at $\alpha = 0.05$.

Results. Table 3 reports per-case results across all $n = 20$ respondents. REMEMORIZE was identified as the less hallucinated summary in 90%, 90%, and 100% of responses for the gastroenterology, cardiology, and general surgery cases, respectively; all three rates exceed chance (dominance $p < 0.001$). Hallucination severity ratings were consistently lower for REMEMORIZE than for Mistral-7B-Instruct across all cases ($r_{\mathrm{rb}} \geq 0.82$, $p \leq 0.001$). Pooling across all 60 respondent–case pairs, 93.3% of responses favored REMEMORIZE ($W = 108.5$, $p < 0.001$, $r_{\mathrm{rb}} = 0.877$).

F Training Details

All experiments are conducted on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB VRAM). We report the complete training configuration for REMEMORIZE, including optimization, memory, and reward parameters, to ensure full reproducibility. We report the complete training configuration for REMEMORIZE, including optimization, memory, and reward parameters, to ensure full reproducibility.

F.1 Data Extraction

All experiments are conducted using publicly available clinical summarization datasets. The MIMIC-

Table 6: Training hyperparameters for REMEMORIZE.

Parameter	Value
Learning rate	1×10^{-5}
Optimizer	AdamW
Clipping radius ϵ_{clip}	0.2
Rollout group size G	4
KL early-stop threshold δ_{KL}	0.02
GRU memory dimension d_m	128
Key/value dimension (d_k, d_v)	(64, 64)
Gate bias init	-2.0
Reward weights $(\alpha, \beta, \gamma, \delta)$	(1.0, 0.3, 0.1, 0.8)
λ_{mem}	0.1

Table 7: Ablation study on MIMIC-IV (n=500). MC = MiniCheck-Flan-T5-Large factuality score (higher↑ is better). The backbone row uses the unmodified `LLaMA-3.1-8B-Instruct`; its higher ROUGE reflects extractive copying rather than improved abstraction. Values are reported as mean ± bootstrap standard deviation.

Variant	Rouge-1	Rouge-L	Factuality↑
w/o GRPO	$19.5_{\pm 0.3}$	$9.5_{\pm 0.2}$	$50.8_{\pm 0.8}$
w/o RL	$19.2_{\pm 0.4}$	$9.4_{\pm 0.2}$	$51.0_{\pm 0.8}$
w/o Memory	$19.3_{\pm 0.3}$	$9.5_{\pm 0.2}$	$51.7_{\pm 0.8}$
w/o Phase-2 update	$21.4_{\pm 0.4}$	$10.8_{\pm 0.3}$	$54.2_{\pm 0.9}$
w/o Adaptive gate	$21.5_{\pm 0.4}$	$10.9_{\pm 0.3}$	$55.6_{\pm 0.9}$
Backbone LLM	$25.5_{\pm 0.5}$	$14.8_{\pm 0.3}$	$54.0_{\pm 1.0}$
Full model	$\mathbf{31.1}_{\pm 0.4}$	$\mathbf{18.8}_{\pm 0.3}$	$\mathbf{74.3}_{\pm 0.9}$

IV-Ext-BHC benchmark (Aali et al., 2024) is derived from the MIMIC-IV electronic health record database and requires credentialed access through PhysioNet, including completion of the associated data use agreement and human-subjects research training. The dataset contains de-identified clinical notes paired with brief hospital course summaries for long-context clinical summarization. The SOAP dataset (Neupane et al., 2024) is publicly available through HuggingFace and consists of patient–provider conversations annotated with structured SOAP-style summaries. All datasets used in this work are fully de-identified and publicly accessible under their respective usage agreements. No private institutional records, proprietary clinical data, or locally collected patient information were used in any stage of training or evaluation.

G Ablation Details

We conduct a series of targeted ablation experiments to isolate the contributions of memory architecture and reward-guided optimization to factual consistency. Each ablation removes or modifies a single component of REMEMORIZE while keeping the remaining training and decoding configuration unchanged. The goal is to determine whether the observed gains arise from adaptive memory control itself, the reinforcement learning objective, or their interaction during autoregressive generation.

Table 7 summarizes the ablation results on the MIMIC-IV-Ext-BHC test set. Across nearly all variants, removing adaptive memory components or reward-guided optimization leads to substantial degradation in factuality, indicating that both memory regulation and policy optimization contribute meaningfully to source-grounded generation.

Training variants. To evaluate the contribution of reward-guided optimization, we modify the reinforcement learning component while preserving the underlying memory architecture. The **w/o GRPO** variant replaces Group Relative Policy Optimization (GRPO) with standard reinforcement learning using global advantage normalization, removing the within-group relative reward estimation used in the full model. The **w/o RL** variant removes reinforcement learning entirely and optimizes the model using only the supervised summarization objective. These variants isolate the extent to which sequence-level reward optimization contributes to factual consistency beyond supervised learning alone.

Both variants substantially reduce factuality relative to the full model, suggesting that adaptive memory control benefits from direct optimization toward factual alignment and hallucination suppression objectives rather than purely token-level likelihood training.

Backbone baseline. We additionally evaluate the unmodified backbone model (`LLaMA-3.1-8B-Instruct`) without memory augmentation or reward-guided training. This baseline serves as a lower bound for assessing the contribution of the proposed framework beyond the pretrained backbone itself. Although the backbone achieves comparatively stronger ROUGE scores than several ablation variants, qualitative inspection indicates that these gains are largely driven by extractive copying behavior rather than improved abstraction or factual grounding. Its lower factuality score further suggests that lexical overlap alone is insufficient to capture

source-faithful clinical generation.

Memory architecture variants. We further investigate the role of adaptive memory control through inference-time modifications applied to the same trained checkpoint. The **w/o Memory** variant removes memory injection entirely, reducing the framework to standard autoregressive decoding without external memory interaction. The **w/o Adaptive Gate** variant replaces the learned gating function with a fixed constant value $\gamma_t = 0.5$, removing dynamic control over information retention and overwrite behavior. The **w/o Phase-2 Update** variant disables memory updates during decoding while retaining the encoded source memory snapshot, preventing the memory state from evolving as generation progresses.

These variants evaluate whether performance depends on the existence of the memory pathway itself, adaptive gating behavior, and continual memory evolution during decoding. The resulting degradation across all three settings suggests that factual consistency depends not only on access to a memory representation, but also on adaptive regulation of how source evidence is preserved throughout autoregressive generation.

H Backbone Generalizability

To evaluate whether the factuality improvements of REMEMORIZE generalize across architectures rather than depending on a specific pretrained backbone, we train and evaluate a second instance of the framework using `Mistral-7B` in place of `Llama-3.1-8B-Instruct`. All remaining components, including the GRU-based memory module, gating network, reward formulation, and GRPO optimization procedure, are kept identical across experiments. This setting isolates the contribution of the proposed memory and policy-learning mechanisms independently of backbone-specific architectural or pretraining effects. Table 4 reports the resulting performance alongside the corresponding backbone-only baselines on both MIMIC-IV-Ext-BHC and SOAP.

Across both datasets, REMEMORIZE consistently improves factual consistency relative to the underlying backbone models. When applied to `Mistral-7B`, the framework improves factuality by +13.1 points on MIMIC-IV-Ext-BHC (51.2 → 64.3) and +8.4 points on SOAP (61.4 → 69.8). On MIMIC-IV-Ext-BHC, lexical and semantic similarity metrics improve alongside factuality; on SOAP, R-L decreases slightly (25.7 → 19.9) while factuality and R-1 improve, consistent with the trade-off observed in the main results. When applied to `Llama-3.1-8B-Instruct`, factuality improves by +15.5 on MIMIC-IV-Ext-BHC and +16.5 on SOAP. On SOAP in particular, ROUGE scores decrease substantially (R-1: 44.3 → 30.7; R-L: 31.6 → 17.9) while factuality improves markedly, reflecting the same pattern whereby reward-guided training prioritizes source grounding over surface-level lexical overlap. This consistency across backbones suggests that the observed improvements are not tied to a particular pretrained model family, tokenizer, or instruction-tuning strategy, and that the gains arise from the memory update mechanism and reward-guided optimization rather than backbone-specific properties.

Overall, these results support the hypothesis that adaptive control over information retention provides a backbone-agnostic mechanism for improving factual grounding during autoregressive generation. The consistency of the improvements across distinct transformer families indicates that REMEMORIZE functions as a lightweight external augmentation rather than a backbone-specific architectural modification.

I MiniCheck Factuality Scoring

Lexical overlap metrics such as ROUGE are not designed to measure factual correctness. A fluent yet hallucinated summary may still achieve high ROUGE scores if it shares surface-level tokens with the reference summary. To directly evaluate source grounding, we adopt MiniCheck (Tang et al., 2024), a fact-verification model fine-tuned from Flan-T5-Large on grounded claim-verification datasets. MiniCheck operates on (document, claim) pairs and evaluates whether a generated claim is supported by the provided source evidence without relying on external knowledge.

Claim decomposition and scoring. Given a source document D and a generated summary $\hat{y}$, we segment $\hat{y}$ into a set of K atomic propositions $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ using sentence boundaries. For each proposition c_k, MiniCheck computes a support probability:

$$p_k = \Pr\big[\textsc{Supported} \mid D,\, c_k\big] \in [0, 1], \quad (14)$$

where $p_k > 0.5$ indicates that c_k is supported by D. The per-summary factuality score is defined as

the proportion of supported propositions:

$$\mathrm{MC}(\hat{y}, D) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\big[p_k > 0.5\big]. \tag{15}$$

Corpus-level factuality is computed as the mean across all N evaluation examples:

$$\overline{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{MC}(\hat{y}^{(i)}, D^{(i)}) \in [0, 1]. \tag{16}$$

Higher $\overline{\mathrm{MC}}$ values indicate that a larger proportion of generated claims are grounded in the source document, with $\overline{\mathrm{MC}} = 1.0$ corresponding to complete source support across all propositions.

Advantages over NLI-based scoring. Unlike standard NLI classifiers such as DeBERTa-based contradiction detectors, MiniCheck is trained specifically for grounded fact verification rather than general natural language inference. This distinction is particularly important in clinical summarization, where strong semantic similarity between sentences can cause NLI models to conflate topical coherence with factual support. By conditioning verification directly on the source document, MiniCheck provides a more reliable estimate of source-grounded factuality in long-form clinical generation.

Implementation. We use the `MiniCheck-Flan-T5-Large` checkpoint released by Tang et al. (2024), which achieves strong performance on the TRUE benchmark (Honovich et al., 2022) among open-weight factuality evaluation models of comparable scale. Inference is performed using a single forward pass per $(\text{document}, \text{claim})$ pair with greedy decoding. Bootstrap confidence intervals (1000 resamples) are reported alongside all MiniCheck scores.

J Reward Weight Sensitivity Analysis

To assess the sensitivity of REMEMORIZE to the reward-weight hyperparameters, we conduct a coordinate-wise grid search over α and δ, corresponding to the factuality reward coefficient and hallucination penalty respectively. These two terms directly govern the primary optimization objectives of the framework. The remaining coefficients, β for coherence and γ for completeness, are treated as auxiliary regularizers and are fixed throughout all experiments at their default values of 0.3 and 0.1 respectively. Because reward configurations are not assumed to transfer across clinical domains, all searches are performed independently on MIMIC-IV-Ext-BHC and SOAP.

The search procedure consists of two sequential sweeps per dataset. In the first sweep, δ is fixed at 0.8 while $\alpha \in \{0.5, 0.7, 1.0, 1.3, 1.6\}$ is evaluated. In the second sweep, α is fixed to the best-performing value from the first stage and $\delta \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ is evaluated. Each configuration is trained for one epoch on 300 samples under otherwise identical optimization settings. Model selection is performed using MiniCheck factuality on the corresponding validation split. Figures 2 and 3 summarize the resulting validation factuality scores with bootstrap standard deviations.

MIMIC-IV-Ext-BHC. On MIMIC-IV-Ext-BHC, factuality peaks at $\alpha^* = 0.7$ with Fact $= 71.23 \pm 1.67$, slightly outperforming the default setting $\alpha = 1.0$ (Fact $= 70.18 \pm 1.69$). Lower values such as $\alpha = 0.5$ noticeably reduce performance, while larger values beyond 1.0 yield no consistent improvement. Overall, the model remains relatively stable across the range $[0.7, 1.6]$, suggesting that the reward formulation is not overly sensitive to moderate variation in the factuality coefficient.

Fixing $\alpha = 0.7$, the hallucination-penalty sweep shows comparatively narrow variation across δ. Performance ranges from 65.32 ± 1.78 at $\delta = 1.0$ to 68.93 ± 1.70 at $\delta = 1.2$, while the default setting $\delta = 0.8$ achieves 68.39 ± 1.65. No monotonic trend emerges across the sweep, indicating that the hallucination penalty does not behave as a simple scaling regularizer. Both weak and intermediate penalties can underperform relative to stronger settings. Taken together, these results suggest that REMEMORIZE remains robust to moderate perturbations around the default configuration $(\alpha = 1.0, \delta = 0.8)$ and does not require highly precise hyperparameter tuning to achieve strong factuality performance.

SOAP. On SOAP, the factuality sweep over α exhibits a different trend. The default configuration $\alpha^* = 1.0$ achieves the strongest validation factuality (Fact $= 49.38$), while both smaller and larger coefficients reduce performance. In contrast to MIMIC-IV-Ext-BHC, the MIMIC-optimal value $\alpha = 0.7$ does not transfer effectively to SOAP, suggesting dataset-specific sensitivity to the factuality reward.

For the hallucination penalty sweep, the default

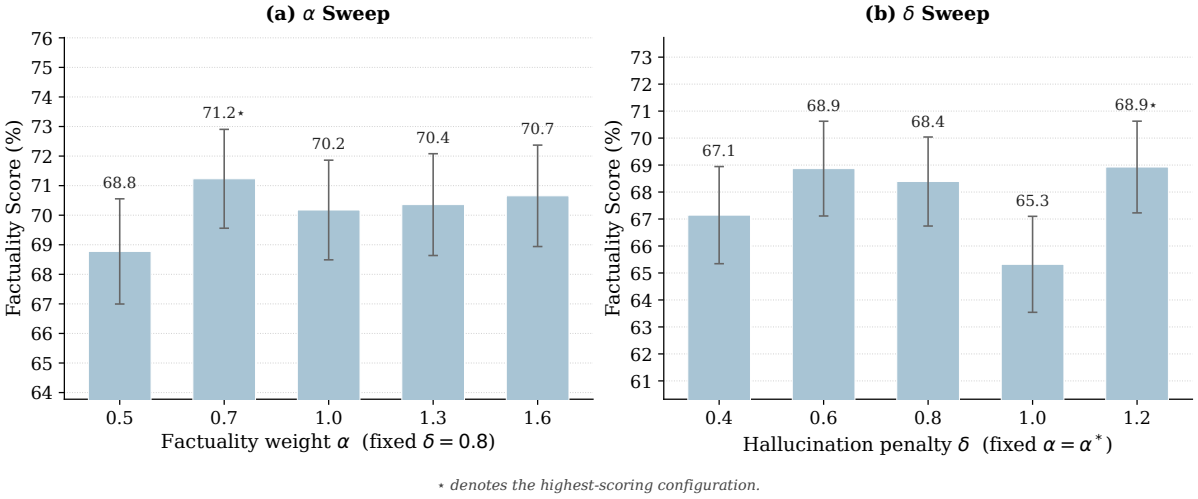

Figure 2: Coordinate-wise reward-weight sensitivity on the MIMIC-IV-Ext-BHC validation set. Left: factuality across five values of α with δ fixed at 0.8. Right: factuality across five values of δ with α fixed at the selected α^*. The best-performing configuration in each sweep is marked with $\star$. Error bars denote bootstrap standard deviation over 1,000 resampling iterations.

setting $\delta^* = 0.8$ again achieves the strongest validation performance (Fact $= 52.71$). Increasing the penalty beyond this point consistently degrades factuality, while reducing it also leads to weaker performance. Unlike MIMIC-IV-Ext-BHC, where several nearby configurations remain competitive, SOAP exhibits a sharper preference for the default setting. This divergence across datasets further motivates independent reward calibration for different clinical summarization domains.

K Memory Gate Activation Analysis

To better understand the behavior of the learned gating policy, we analyze the gate activations, encoded memory statistics, and auxiliary reconstruction objective across several ablation conditions on the MIMIC-IV-Ext-BHC test set. Specifically, we report the average gate activation $\bar{\gamma}$, the ℓ_2 norm of the final encoded memory state $\|\mathbf{m}_{T_{\text{src}}}\|_2$, and the auxiliary reconstruction loss $\mathcal{L}_{\text{aux}}$ introduced during training. These diagnostics provide insight into how reward-guided optimization influences memory retention, overwrite behavior, and the internal consistency of the learned memory representation.

Selective retention behavior. Across all trained variants, the learned gate converges to consistently low activation values, with $\bar{\gamma}$ remaining within the narrow range $[0.09, 0.10]$. This behavior indicates that the model preserves the existing memory state for the majority of source-token updates and commits new information only selectively. Under the GRU-style update mechanism described in Eq. 1, low gate activations correspond to gradual accumulation of information rather than aggressive overwriting of previously stored evidence.

Table 8: Memory gate statistics under ablation conditions on MIMIC-IV-Ext-BHC. $\bar{\gamma}$ denotes the mean gate activation averaged across source tokens and evaluation examples. $\|\mathbf{m}\|_2$ denotes the mean ℓ_2 norm of the final encoded memory state. $\mathcal{L}_{\text{aux}}$ denotes the mean reconstruction MSE between token projections and the final memory representation. Lower $\mathcal{L}_{\text{aux}}$ indicates stronger alignment between local token representations and the global encoded memory state.

Variant	$\bar{\gamma}$	$\|\mathbf{m}\|_2$	$\mathcal{L}_{\text{aux}}$
Full model	0.097	9.29	0.093
w/o Adaptive gate (γ=0.5)	0.500	6.72	0.068
w/o GRPO	0.098	5.43	0.271
w/o RL	0.091	5.37	0.271

This pattern is consistent with the structure of long clinical documents, where only a relatively small subset of tokens contains clinically decisive information. Instead of rewriting memory at every token position, the gating policy learns to preserve stable contextual evidence while updating memory primarily around salient clinical content. The resulting behavior resembles sparse evidence integration rather than uniform token-level averaging.

Impact of reward-guided optimization. The strongest contrast in Table 8 emerges between the full model and the variants trained without rein-

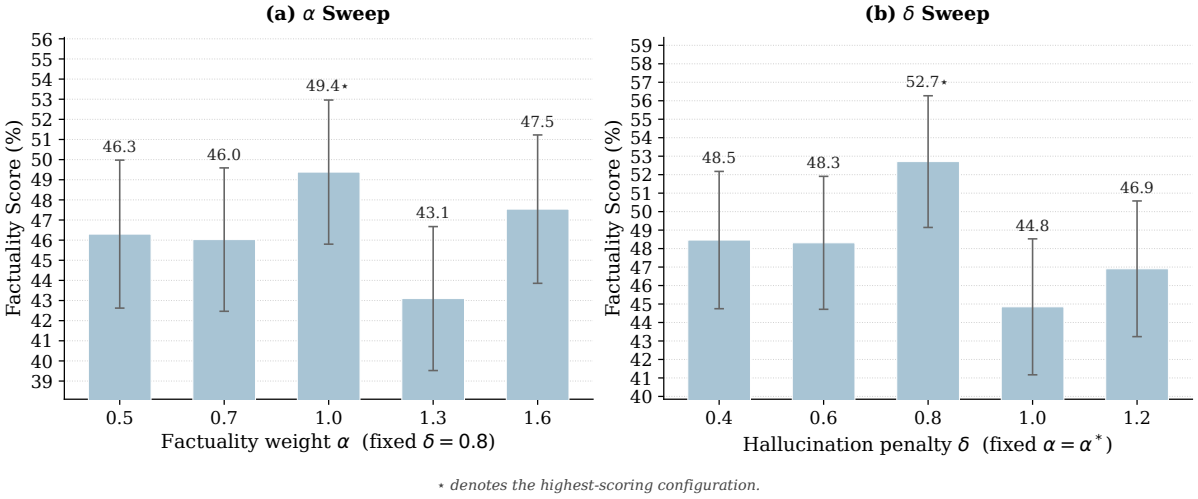

Figure 3: Coordinate-wise reward-weight sensitivity on the SOAP validation set. Left: factuality across five values of α with δ fixed at 0.8. Right: factuality across five values of δ with α fixed at $\alpha^* = 1.0$. The best-performing configuration in each sweep is marked with $\star$. Error bars denote bootstrap standard deviation over 1,000 resampling iterations.

forcement learning. Although the average gate activation remains nearly unchanged across the full model, w/o RL, and w/o GRPO conditions, the encoded memory statistics differ substantially. Removing reinforcement learning reduces the memory norm from 9.29 to approximately 5.4, suggesting that reward-guided optimization encourages the memory state to accumulate richer and more expressive representations of the source document.

An even larger difference appears in the auxiliary reconstruction loss. When reinforcement learning is removed, $\mathcal{L}_{\text{aux}}$ increases from 0.093 to 0.271, indicating substantially weaker alignment between individual token projections and the final encoded memory state. This result suggests that the reward signal influences not only the gating policy itself, but also the overall coherence of the learned memory representation. In other words, reward-guided optimization appears to shape both *when* information is retained and *how* source evidence is internally organized within memory.

Fixed-gate behavior. The fixed-gate variant, where $\gamma_t = 0.5$ for all time steps, produces qualitatively different memory dynamics. Because retention and overwrite operations receive equal weighting throughout the sequence, the resulting memory trajectory behaves more like a smoothed running average over token representations. This produces a lower memory norm (6.72) and a lower reconstruction loss (0.068) relative to the full model.

Although the reduced reconstruction loss may initially appear favorable, it reflects a substantially less selective memory mechanism. A fixed midpoint gate encourages uniform averaging across tokens rather than prioritization of clinically relevant evidence. This interpretation is consistent with the factuality degradation observed in the main ablation study, where the fixed-gate variant underperforms the adaptive policy despite achieving lower reconstruction error. Taken together, these findings suggest that effective long-context clinical summarization requires not only stable memory encoding, but also adaptive regulation over which information should be retained during generation.

L Illustrative Hallucination Case Study

To illustrate the type of hallucination targeted by REMEMORIZE, we present a representative example from the SOAP (MTS-Dialog) test set in which a baseline language model generates unsupported clinical content from a minimal source conversation. This qualitative example complements the quantitative factuality metrics reported in Table 1.

Source Conversation. The following exchange constitutes the full clinical interaction provided to the model:

> *Doctor:* How are you feeling now after the treatment?
> *Patient:* I am feeling pretty fair, improving slowly.
> *Doctor:* Okay.

Reference Summary. The reference SOAP note reflects only the information explicitly stated in the conversation:

> **Subjective:** Patient feels "pretty fair" and reports improving slowly.
> **Objective, Assessment, Plan:** No data provided in this encounter.

Baseline Prediction. A representative zero-shot baseline instead generates the following summary:

> Patient presented with progressive shortness of breath over the past three months, with worsening symptoms during the last two weeks. The patient reported a history of cigarette smoking, which she discontinued one year ago, as well as daily alcohol consumption. Current medications included aspirin, vitamin D3, calcium supplementation, omega-3 fatty acids, atenolol, lisinopril, and pravastatin. Family history was notable for chronic obstructive pulmonary disease, hypertension, and abdominal aortic aneurysm. Physical examination demonstrated bilateral wheezing, decreased air entry, mild crackles, elevated jugular venous pressure, tachycardia, pallor, and bilateral lower extremity edema.

The generated summary introduces numerous unsupported clinical details, including respiratory symptoms, smoking history, medication usage, family history, and abnormal examination findings that do not appear in the source conversation. Although the output is fluent and clinically plausible, none of these claims are grounded in the provided evidence. This example illustrates a common failure mode in long-form clinical generation, where the model produces medically coherent narratives by relying on pretrained statistical associations rather than source-supported evidence. The example further motivates the need for mechanisms that explicitly regulate how source information is retained throughout autoregressive generation.