REMEMORIZE: REWARD-GUIDED ADAPTIVE MEMORY MANAGEMENT FOR FAITHFUL CLINICAL TEXT SUMMARIZATION

Anonymous authors
Paper under double-blind review

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 with gated updates during decoding. The interface is trained using supervised memory-consistency objectives, group-relative reward-weighted candidate learning, and a clipped score-weighted auxiliary gate objective based on sequence-level signals for factual alignment, coherence, completeness, and hallucination suppression. Across two popular summarization benchmarks, `ReMemorize` improves factual consistency over strong baselines, with gains concentrated on faithfulness rather than surface-level lexical overlap. Our results suggest that explicit control over information retention can improve factual grounding in clinical summarization. Our code is publicly 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 (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), retrieval augmentation (Borgeaud et al., 2022; Ram et al., 2023; Piya & Beheshti, 2025), 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).

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, but generally focus on what information is available rather than token-level updates to a dedicated recurrent memory state. In standard autoregressive transformers, information retention depends on hidden-state transitions and attention mechanisms fixed by the architecture (Khandelwal et al., 2018; Sun et al., 2021; Mallen et al., 2023). We hypothesize that explicitly regulating a separate memory state during decoding could help preserve source evidence and reduce factual drift. This question is particularly relevant to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

clinical summarization, where unsupported or contradicted statements can affect the reliability of generated documentation.

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., 2019). 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 reward-guided and preference optimization approaches (Ouyang et al., 2022; Rafailov et al., 2023), we formulate gate selectivity as a reward-guided learning problem, combining group-relative candidate evaluation (Shao et al., 2024) with reward-weighted memory training and a clipped score-weighted auxiliary gate objective. During inference, the gate parameters remain fixed, while the memory state is updated 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 a recurrent memory interface that adaptively regulates evidence retention during source encoding and autoregressive decoding over a frozen language model.
- We develop a reward-guided training framework combining group-relative candidate evaluation, reward-weighted teacher-forced learning, an auxiliary memory-consistency objective, and a clipped score-weighted gate objective.
- We demonstrate improved factual consistency on two clinical summarization benchmarks using automatic, model-based, and human evaluation.

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 & Beheshti, 2026; Yuan et al., 2025). Guided continued pre-training on clinical corpora has separately been shown to improve factual reliability (Elhady et al., 2024); such methods modify model parameters prior to inference, whereas REMEMORIZE instead introduces a decoding-time memory mechanism that requires no additional pretraining stage. Other methods address long clinical inputs through budgeted context selection and routing (Qureshi et al., 2026), or ground longitudinal clinical records using temporal and guideline knowledge graphs (Li et al., 2026). Despite these advances, maintaining factual consistency in clinical summaries remains challenging. Context selection determines which source evidence is presented to the generator, while external grounding can supply additional evidence. Neither, by itself, governs how a separate memory state retains information during autoregressive decoding. We therefore study gated recurrent updates to that state and evaluate their effect on factual consistency across two clinical summarization benchmarks.

Factual Consistency under Long-Context Generation. Factual consistency between generated summaries and source clinical records remains challenging, as unsupported statements can affect downstream clinical decision-making. LLMs may generate fluent yet ungrounded content even when relevant evidence is present in the input (Ji et al., 2023; Maynez et al., 2020; McKenna et al., 2023). Studies of long-context use further show that models do not always make effective use of evidence throughout an input document (Liu et al., 2024). Existing work examines hallucination detection through uncertainty estimation (Farquhar et al., 2024) and mitigation through verification-based methods (Liu et al., 2025). Motivated by these challenges, REMEMORIZE incorporates a separate recurrent memory state with gated updates during clinical summary generation.

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., 2019) and Compressive Transformer (Rae et al., 2019) incorporate recurrent memory, while models including Memformer (Wu et al., 2022), MART (Lei et al., 2020), and gated linear-attention variants such as GLA (Yang et al., 2023) employ memory updates within sequence modeling architectures. Mamba (Gu & Dao, 2023) and Gated DeltaNet (Yang et al., 2025b) further demonstrate input-dependent state updates, so gated memory updates themselves are well established. In summarization, Sliding Selector Networks (Cui & Hu, 2021), NeuSum (Zhou et al., 2018), and MemSum (Gu

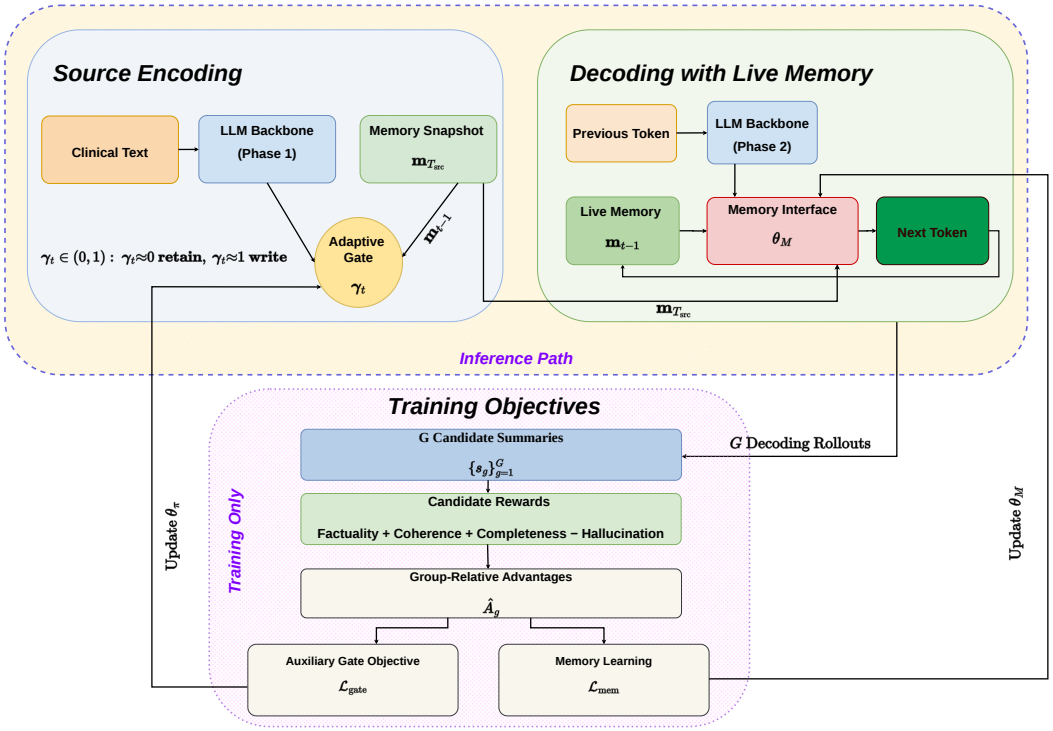

Figure 1: Overview of the REMEMORIZE framework. During training, the model generates provisional summaries evaluated using a composite reward function. During inference, the policy is frozen.

et al., 2022) use memory to track extraction history and reduce redundancy; memory-aware methods have also been studied for long dialogues (Wang et al., 2025). We examine whether a recurrent memory interface coupled to a frozen language model can improve factual consistency in clinical summarization when its state continues to update during generation.

Reward-Guided and Preference-Based Summarization. Reward-guided and preference-based methods have been used to align language models with objectives such as relevance, coherence, and factual consistency. Direct Preference Optimization (Rafailov et al., 2023), RRHF (Yuan et al., 2023), and RLAIF (Lee et al., 2024) use preference feedback to guide generation, while self-rewarding approaches support iterative refinement (Yuan et al., 2024). In summarization, such methods target properties that are difficult to capture through token-level likelihood objectives alone (Ryu et al., 2024; Li et al., 2025; Song et al., 2026; Yadav et al., 2025; Yang et al., 2025a). Textual-entailment rewards have been used to improve factual consistency in summarization (Roit et al., 2023). Related work also explores agentic frameworks that decompose generation into drafting, critique, and revision stages (Tu et al., 2025; Piya & Beheshti, 2026). REMEMORIZE applies candidate-level reward signals to train a memory interface coupled to a frozen language model, with gated memory updates continuing during generation.

3 METHODOLOGY

REMEMORIZE addresses factual inconsistency in clinical summarization by maintaining an explicit memory state during autoregressive generation. To address this, the framework maintains a latent memory state $\mathbf{m}_t$ that is initialized during source encoding and evolves through gated updates during decoding. As shown in Figure 1, the framework consists of three components: *Source Encoding* (Phase 1), *Decoding with Live Memory* (Phase 2), and *Reward-Guided Training*, which

computes rewards over the completed rollouts from both phases. The full procedure is summarized in Algorithm 1 (Appendix A).

Problem Formulation. Given an input document d, the model generates a summary s^* while maintaining a latent memory state $\mathbf{m}_t$ throughout source encoding and 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 $t \in \{1, \ldots, T_{\text{src}}\}$, where T_{src} is the number of source tokens. 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., 2019), 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 single-gate recurrent update inspired by GRU update mechanisms (Cho 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.

The gate uses a two-layer MLP conditioned on the current token key $k_t \in \mathbb{R}^{d_k}$ and 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 to the key dimension, and the MLP outputs a scalar logit. Joint conditioning enables the gate to limit redundant writes and prioritize new information. Its output bias is initialized to -2.0 ($\sigma(-2) \approx 0.12$), keeping the gate near-closed early in training. The continuous γ_t interpolates between retaining the previous state and writing the current token representation. In Phase 1, the source is processed token by token in a causal forward pass (Rae et al., 2019; Wu et al., 2022); at each $t \in \{1, \ldots, T_{\text{src}}\}$, Eq. 2 determines the update in Eq. 1. The final state $\mathbf{m}_{T_{\text{src}}}$ is a fixed-size representation of the source.

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 single-gate update in Eqs. 1 and 2, advancing the memory to $\mathbf{m}_{t'}$ before the next decoding step.

3.3 Reward-Guided Training

Training targets two components: the gate network θ_π and the memory-interface parameters θ_M (value, memory, and write projections; key_proj is held at initialization throughout, as it receives no

gradient under this objective); the LLM backbone remains frozen throughout. Rewards are defined over complete generation trajectories relative to the source document. Factual summarization requires reference-free rewards that remain sensitive to hallucination. Prior work has used natural language inference (NLI) entailment probability as a 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:

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

where $\alpha, \beta, \gamma, \delta > 0$ are scalar weights (Details in Appendix F).

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 cross-encoder fine-tuned for natural language inference (He et al., 2021). Because such cross-encoders accept a bounded number of tokens per pair, source documents exceeding this length are partitioned into overlapping windows, each scored independently against s^*; R_{factual} and $R_{\text{hallucination}}$ take the maximum entailment and contradiction probability, respectively, over all windows (implementation details in Appendix I). 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).

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 reward-conditioned optimization across documents.

Group-Relative Advantage Estimation. For each document d, the memory-conditioned model generates a group of G candidate summaries $\{s_g\}_{g=1}^{G}$. Each candidate s_g receives the composite reward $R(s_g)$ from Eq. 4, i.e. the same factual alignment, hallucination suppression, coherence, and completeness terms evaluated on s_g. Following the group-relative normalization introduced by Shao et al. (2024), we compute

$$\hat{A}_g = \frac{R(s_g) - \mu_G}{\sigma_G + \epsilon_0}, \tag{7}$$

where $\mu_G = \frac{1}{G}\sum_{g=1}^{G} R(s_g)$, $\sigma_G = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(R(s_g) - \mu_G)^2}$, and ϵ_0 is a numerical stabilizer.

The resulting advantage $\hat{A}_g$ is the single training signal that drives both trainable components identified at the start of this section: it supplies a reward-weighted supervision target for the memory-interface parameters θ_M (Reward-Weighted Candidate Learning, below) and a score-weighted signal for the gate network θ_π (Auxiliary Gate Objective, below); together these parameters govern the memory update and gating computations used in both Phase 1 and Phase 2 (Eqs. 1–2, Figure 1).

Reward-Weighted Candidate Learning. Let $g^\star = \arg\max_g R(s_g)$ denote the highest-scoring candidate. When $\hat{A}_{g^\star} > 0$, this candidate provides an additional teacher-forced training target, converting sequence-level feedback into a differentiable learning signal for the memory-interface parameters θ_M:

$$\mathcal{L}_{\text{mem}} = \lambda_{\text{LM}} \Big[\mathcal{L}_{\text{CE}}^{\text{gold}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}(s^{\text{gold}}) + \lambda_{\text{rwt}}\, [\hat{A}_{g^\star}]_+ \big(\mathcal{L}_{\text{CE}}^{\text{cand}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}(s^*_{g^\star}) \big) \Big], \tag{8}$$

where $\mathcal{L}_{\text{CE}}^{\text{gold}}$ and $\mathcal{L}_{\text{CE}}^{\text{cand}}$ are cross-entropy losses against the gold reference and $s^*_{g^\star}$ respectively, $[\hat{A}_{g^\star}]_+ = \max(\hat{A}_{g^\star}, 0)$, and $\lambda_{\text{LM}}, \lambda_{\text{aux}}, \lambda_{\text{rwt}}$ are fixed coefficients (Appendix F). For either target text $x \in \{s^{\text{gold}}, s^*_{g^\star}\}$:

$$\mathcal{L}_{\text{aux}}(x) = \frac{1}{T_x\, d_m} \sum_{t'=1}^{T_x} \left\| \tanh\Big(\text{write_proj}(v_{t'}^{(x)})\Big) - \text{sg}\Big(\mathbf{m}_{T_x}^{(x)}\Big) \right\|_2^2, \tag{9}$$

where T_x is the number of decoding steps for x and $\text{sg}(\cdot)$ denotes stop-gradient; this objective encourages decoding-time write projections to remain compatible with the memory state produced for the corresponding target sequence. θ_M is updated via AdamW (Loshchilov & Hutter, 2017) using $\mathcal{L}_{\text{mem}}$.

Auxiliary Gate Objective. In parallel with reward-weighted memory learning, the gate network is updated using a group-relative clipped objective. For each candidate index g, binary variables are sampled from the gate probabilities over the source key window,

$$a_{g,t} \sim \text{Bernoulli}(\gamma_t^{\text{old}}), \qquad t = 1, \ldots, K, \tag{10}$$

where K is the number of source-window gate decisions, shared across all G candidates. Their joint log-likelihood is

$$\ell_g(\theta_\pi) = \sum_{t=1}^{K} \left[a_{g,t} \log \gamma_t + (1 - a_{g,t}) \log(1 - \gamma_t) \right], \tag{11}$$

with ratio $\rho_g = \exp\left[\ell_g(\theta_\pi) - \ell_g(\theta_\pi^{\text{old}})\right]$ and objective

$$\mathcal{L}_{\text{gate}} = -\frac{1}{G} \sum_{g=1}^{G} \min\Big(\rho_g \hat{A}_g,\ \text{clip}(\rho_g, 1{-}\epsilon_{\text{clip}}, 1{+}\epsilon_{\text{clip}})\, \hat{A}_g \Big) - \lambda_H\, \mathcal{H}(q_{\theta_\pi}) \tag{12}$$

Equation 12 uses clipped likelihood ratios (Schulman et al., 2017), with clipping radius ϵ_{clip} and entropy coefficient λ_H. Binary gate samples enter only the auxiliary objective; candidate generation and memory updates use continuous γ_t (Eqs. 1–2). Thus, $\hat{A}_g$ is not the return of a sampled binary-gate trajectory, and $\mathcal{L}_{\text{gate}}$ is a heuristic score-weighted objective rather than an unbiased policy-gradient estimator for generated summaries. Recurrent state transitions are treated as non-differentiable forward updates, and θ_π is optimized only through $\mathcal{L}_{\text{gate}}$ (Appendix A, Algorithm 1). Updates stop early when the approximate KL divergence between successive gate distributions exceeds δ_{KL} (Table 6). After Phase 1 encoding and Phase 2 decoding, $\hat{A}_g$ informs supervised memory-interface learning for θ_M (Eqs. 8–9) and the clipped gate update for θ_π (Eqs. 11–12), as summarized in Figure 1.

4 Experimental Setup

Datasets We evaluate on two clinical summarization benchmarks. **MIMIC-IV-Ext-BHC** (Aali et al., 2025) contains 270,033 clinical notes paired with brief hospital course summaries (avg. 2,267 tokens); we use 1,000/500/500 for train/val/test. **SOAP (ClinicSum)** (Neupane et al., 2024) comprises 1,473 patient-doctor conversation-summary pairs assembled from 272 Figshare conversations and 1,201 MTS-Dialog conversations (Abacha et al., 2023), with structured SOAP summaries; we use 1,023/150/300 for train/val/test.

Evaluation Metrics Generated summaries are evaluated using lexical overlap, semantic similarity, source-grounded factuality, LLM-judge ratings, and human ratings of hallucination severity. **Lexical and Semantic.** We report ROUGE (Lin, 2004) and BERTScore F1 (Zhang et al., 2020). **Factuality.** MiniCheck (Tang et al., 2024) assesses whether claims in each generated summary are supported by the source document, yielding the factuality score (Fact.↑; Appendix J). **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=30) compared REMEMORIZE against Mistral-7B across three clinical cases, with participants rating hallucination severity (Appendix E). The prompt used to identify hallucinated spans for qualitative analysis is provided in Appendix B.

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 (↑ higher is better). All values are mean$_{\pm \text{std}}$ over 1000 bootstrap iterations. Bold: best per column. Human-rated hallucination severity is reported separately in Table 3.

	MIMIC-IV-Ext-BHC					SOAP				
Model	**R-1↑**	**R-2↑**	**R-L↑**	**BS-F1↑**	**Fact.↑**	**R-1↑**	**R-2↑**	**R-L↑**	**BS-F1↑**	**Fact.↑**
General-purpose models										
LLaMA-3.2-1B	$23.4_{\pm 0.3}$	$4.2_{\pm 0.2}$	$13.1_{\pm 0.2}$	$80.9_{\pm 0.1}$	$66.6_{\pm 0.8}$	$22.2_{\pm 0.8}$	$5.8_{\pm 0.4}$	$13.2_{\pm 0.5}$	$82.6_{\pm 0.1}$	$65.2_{\pm 1.3}$
LLaMA-3.2-3B	$21.2_{\pm 0.4}$	$3.9_{\pm 0.2}$	$11.7_{\pm 0.2}$	$80.2_{\pm 0.1}$	$52.7_{\pm 0.8}$	$30.3_{\pm 0.5}$	$14.7_{\pm 0.5}$	$25.7_{\pm 0.5}$	$86.0_{\pm 0.1}$	$46.2_{\pm 1.1}$
Qwen2.5-3B	$28.3_{\pm 0.4}$	$6.1_{\pm 0.2}$	$15.9_{\pm 0.2}$	$82.7_{\pm 0.1}$	$62.3_{\pm 0.7}$	$28.0_{\pm 0.6}$	$16.9_{\pm 0.5}$	$26.7_{\pm 0.6}$	$87.3_{\pm 0.1}$	$50.3_{\pm 1.0}$
Mistral-7B	$28.6_{\pm 0.3}$	$6.1_{\pm 0.2}$	$15.8_{\pm 0.2}$	$82.7_{\pm 0.1}$	$51.2_{\pm 0.7}$	$27.5_{\pm 0.5}$	$14.1_{\pm 0.4}$	$25.7_{\pm 0.4}$	$83.6_{\pm 0.1}$	$61.4_{\pm 0.9}$
Qwen2.5-7B	$29.3_{\pm 0.2}$	$7.0_{\pm 0.2}$	$17.2_{\pm 0.2}$	$82.1_{\pm 0.1}$	$58.4_{\pm 0.6}$	$26.0_{\pm 0.6}$	$16.5_{\pm 0.5}$	$25.0_{\pm 0.5}$	$83.8_{\pm 0.1}$	$51.9_{\pm 1.1}$
Phi-3.5-mini	$26.5_{\pm 0.3}$	$5.2_{\pm 0.2}$	$14.9_{\pm 0.2}$	$82.7_{\pm 0.1}$	$53.5_{\pm 0.7}$	$29.9_{\pm 0.5}$	$16.6_{\pm 0.4}$	$\mathbf{28.2}_{\pm 0.4}$	$\mathbf{87.5}_{\pm 0.1}$	$60.4_{\pm 1.0}$
Gemma-2-2B	$28.3_{\pm 0.3}$	$6.6_{\pm 0.2}$	$16.0_{\pm 0.2}$	$82.3_{\pm 0.1}$	$65.1_{\pm 0.7}$	$28.6_{\pm 0.5}$	$9.6_{\pm 0.3}$	$19.2_{\pm 0.3}$	$83.9_{\pm 0.1}$	$60.9_{\pm 1.0}$
Flan-T5-XXL	$24.2_{\pm 0.5}$	$6.4_{\pm 0.4}$	$14.6_{\pm 0.4}$	$83.3_{\pm 0.1}$	$71.9_{\pm 1.0}$	$17.7_{\pm 0.7}$	$8.8_{\pm 0.6}$	$14.9_{\pm 0.7}$	$83.7_{\pm 0.1}$	$71.1_{\pm 2.1}$
Biomedical and clinical domain-adapted models										
BioMistral-7B	$28.1_{\pm 0.4}$	$7.0_{\pm 0.3}$	$17.0_{\pm 0.3}$	$82.6_{\pm 0.1}$	$69.3_{\pm 0.9}$	$26.1_{\pm 0.7}$	$8.1_{\pm 0.6}$	$16.5_{\pm 0.7}$	$82.6_{\pm 0.1}$	$50.9_{\pm 1.8}$
MedAlpaca-7B	$22.3_{\pm 0.4}$	$4.5_{\pm 0.2}$	$13.0_{\pm 0.2}$	$82.4_{\pm 0.1}$	$53.8_{\pm 1.0}$	$26.2_{\pm 0.7}$	$10.3_{\pm 0.5}$	$18.8_{\pm 0.6}$	$83.0_{\pm 0.1}$	$47.2_{\pm 1.3}$
SummOrchestra-8B	$24.6_{\pm 0.4}$	$7.0_{\pm 0.2}$	$16.1_{\pm 0.3}$	$80.9_{\pm 0.1}$	$71.0_{\pm 0.9}$	$\mathbf{33.0}_{\pm 0.7}$	$16.6_{\pm 0.7}$	$26.6_{\pm 0.8}$	$85.8_{\pm 0.1}$	$70.6_{\pm 1.6}$
Reasoning-oriented model										
DeepSeek-R1-8B	$13.5_{\pm 0.5}$	$1.2_{\pm 0.3}$	$8.7_{\pm 0.3}$	$80.6_{\pm 0.2}$	$70.8_{\pm 0.5}$	$10.8_{\pm 0.6}$	$1.1_{\pm 0.3}$	$7.1_{\pm 0.4}$	$80.2_{\pm 0.3}$	$61.3_{\pm 0.8}$
Extractive model										
MemSum-PubMed	$19.0_{\pm 0.5}$	$5.2_{\pm 0.3}$	$11.9_{\pm 0.3}$	$81.0_{\pm 0.3}$	$58.3_{\pm 0.6}$	$18.2_{\pm 0.4}$	$7.8_{\pm 0.5}$	$13.9_{\pm 0.5}$	$82.5_{\pm 0.2}$	$52.2_{\pm 0.9}$
Backbone comparisons										
LLaMA-3.1-8B (backbone)	$27.1_{\pm 0.3}$	$5.1_{\pm 0.2}$	$13.1_{\pm 0.2}$	$82.2_{\pm 0.1}$	$51.5_{\pm 0.7}$	$27.4_{\pm 0.6}$	$8.0_{\pm 0.5}$	$14.9_{\pm 0.6}$	$82.8_{\pm 0.1}$	$46.5_{\pm 1.1}$
LLaMA-3.1-8B (SFT)	$30.1_{\pm 0.4}$	$6.5_{\pm 0.2}$	$14.0_{\pm 0.3}$	$82.4_{\pm 0.2}$	$52.3_{\pm 0.6}$	$23.9_{\pm 0.5}$	$12.9_{\pm 0.4}$	$22.0_{\pm 0.6}$	$86.8_{\pm 0.2}$	$51.6_{\pm 0.2}$
ReMemorize	$\mathbf{31.1}_{\pm 0.3}$	$\mathbf{7.1}_{\pm 0.2}$	$\mathbf{18.8}_{\pm 0.2}$	$\mathbf{83.9}_{\pm 0.1}$	$\mathbf{74.3}_{\pm 1.7}$	$30.7_{\pm 0.5}$	$\mathbf{18.2}_{\pm 0.3}$	$17.9_{\pm 0.3}$	$84.2_{\pm 0.1}$	$\mathbf{71.8}_{\pm 2.3}$

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 (Grattafiori et al., 2024) to assess scaling within a single architecture, along with Qwen2.5 models at 3B and 7B (Qwen Team et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Phi-3.5-mini-Instruct (Abdin et al., 2024), Gemma-2-2B-IT (Team et al., 2024), and Flan-T5-XXL (Chung et al., 2024) as an encoder-decoder baseline, covering instruction-tuned models spanning the 1B–11B parameter range. **Domain-adapted models.** We include BioMistral-7B (Labrak et al., 2024), MedAlpaca-7B (Han et al., 2023), and SummOrchestra-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 (Guo et al., 2025) 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 We use a frozen LLaMA-3.1-8B-Instruct backbone coupled to a gated recurrent memory module, trained with reward-weighted candidate learning and an auxiliary gate objective. For each document, we generate G=4 candidate summaries to compute group-relative advantages. Hyperparameters and reward-weight sensitivity analyses are provided in Appendices F and K.

5 RESULTS

Table 1 reports automatic evaluation results. Across both datasets, REMEMORIZE achieves the highest factuality scores while remaining competitive on summary quality metrics. On MIMIC-IV-Ext-BHC, it achieves a factuality score of 74.3, exceeding the strongest baseline, Flan-T5-XXL (71.9), by 2.4 points. It also attains the highest ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 scores, improving BERTScore F1 from 83.3 for the strongest baseline to 83.9. The supervised LLaMA-3.1-8B baseline trained on the same gold summaries trails REMEMORIZE in both ROUGE-1 (30.1 vs. 31.1) and factuality (52.3 vs. 74.3). On SOAP, REMEMORIZE again achieves the highest factuality score (71.8), exceeding Flan-T5-XXL (71.1) by 0.7 points. It also achieves the highest ROUGE-2 score (18.2), although SummOrchestra-8B leads in ROUGE-1 (33.0 vs. 30.7), Phi-3.5-mini leads in ROUGE-L (28.2 vs. 17.9) and BERTScore F1 (87.5 vs. 84.2). Thus, the SOAP gains in factuality do not coincide with uniformly higher lexical or semantic overlap. 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 from source sentences, it remains below several generative baselines in factuality on both datasets. Overall, REMEMORIZE's gains extend across factuality and summary quality metrics on MIMIC-IV-Ext-BHC, whereas its gains on SOAP are concentrated in factuality and ROUGE-2.

Table 2 reports LLM-as-a-judge results. REMEMORIZE achieves the highest mean factual-consistency score on both datasets (4.68). On MIMIC-IV-Ext-BHC, it also has the lowest hallucination score (1.06 vs. 1.10) and highest coherence score (4.83 vs. 4.82). Its completeness score exceeds all generative baselines (4.29 vs. 4.21) but trails the extractive MemSum-PubMed model (4.50). On SOAP, it leads in completeness (4.32 vs. 4.28) and coherence (4.86 vs. 4.83), while its hallucination score trails the best baseline (1.30 vs. 1.03). SummOrchestra-8B scores 4.41 and 4.47 in factual consistency on MIMIC-IV-Ext-BHC and SOAP, respectively. Appendix M provides a qualitative hallucination case study.

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, 93.3% in 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.

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	**Halluci-nation↓**	**Factual Consist.↑**	**Complete-ness↑**	**Coher-ence↑**	**Halluci-nation↓**	**Factual Consist.↑**	**Complete-ness↑**	**Coher-ence↑**
General-purpose models								
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	$1.14_{\pm 0.39}$	$4.60_{\pm 0.23}$	$4.21_{\pm 0.39}$	$4.82_{\pm 0.25}$	$1.06_{\pm 0.21}$	$4.59_{\pm 0.41}$	$4.12_{\pm 0.57}$	$4.81_{\pm 0.35}$
Mistral-7B	$1.45_{\pm 0.53}$	$4.48_{\pm 0.22}$	$4.02_{\pm 0.36}$	$4.81_{\pm 0.26}$	$1.65_{\pm 0.25}$	$4.55_{\pm 0.26}$	$4.09_{\pm 0.44}$	$4.83_{\pm 0.26}$
Qwen2.5-7B	$1.21_{\pm 0.46}$	$4.66_{\pm 0.26}$	$4.14_{\pm 0.37}$	$4.40_{\pm 0.21}$	$1.35_{\pm 0.52}$	$4.55_{\pm 0.33}$	$4.28_{\pm 0.43}$	$4.67_{\pm 0.24}$
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	$1.71_{\pm 0.81}$	$4.57_{\pm 0.49}$	$3.52_{\pm 0.84}$	$4.78_{\pm 0.38}$	$1.71_{\pm 0.70}$	$3.50_{\pm 1.21}$	$2.58_{\pm 0.97}$	$4.11_{\pm 1.08}$
Biomedical and clinical domain-adapted models								
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	$1.32_{\pm 0.60}$	$4.26_{\pm 0.71}$	$3.44_{\pm 0.83}$	$4.52_{\pm 0.67}$	$1.48_{\pm 0.63}$	$4.34_{\pm 0.75}$	$3.85_{\pm 0.92}$	$4.56_{\pm 0.63}$
SummOrchestra-8B	$1.36_{\pm 0.58}$	$4.41_{\pm 0.48}$	$3.33_{\pm 0.63}$	$4.76_{\pm 0.43}$	$1.60_{\pm 0.62}$	$4.47_{\pm 0.46}$	$3.82_{\pm 0.69}$	$4.81_{\pm 0.32}$
Reasoning-oriented model								
DeepSeek-R1-8B	$1.58_{\pm 0.65}$	$4.29_{\pm 0.80}$	$4.15_{\pm 0.86}$	$4.62_{\pm 0.77}$	$1.63_{\pm 0.69}$	$4.31_{\pm 0.70}$	$4.19_{\pm 0.80}$	$4.65_{\pm 0.67}$
Extractive model								
MemSum-PubMed	$2.45_{\pm 0.55}$	$3.50_{\pm 0.74}$	$\mathbf{4.50}_{\pm 0.32}$	$4.11_{\pm 0.65}$	$1.67_{\pm 0.65}$	$3.14_{\pm 1.32}$	$2.33_{\pm 0.98}$	$3.72_{\pm 1.23}$
Backbone comparisons								
LLaMA-3.1-8B (backbone)	$1.22_{\pm 0.46}$	$4.55_{\pm 0.21}$	$4.04_{\pm 0.40}$	$4.82_{\pm 0.25}$	$\mathbf{1.03}_{\pm \mathbf{0.17}}$	$4.53_{\pm 0.49}$	$3.79_{\pm 0.75}$	$4.76_{\pm 0.38}$
LLaMA-3.1-8B (SFT)	$1.19_{\pm 0.51}$	$3.67_{\pm 1.23}$	$3.02_{\pm 1.05}$	$4.05_{\pm 1.21}$	$1.20_{\pm 0.43}$	$4.47_{\pm 0.47}$	$3.77_{\pm 0.82}$	$4.63_{\pm 0.46}$
ReMemorize	$\mathbf{1.06}_{\pm \mathbf{0.23}}$	$\mathbf{4.68}_{\pm \mathbf{0.17}}$	$4.29_{\pm 0.14}$	$\mathbf{4.83}_{\pm \mathbf{0.14}}$	$1.30_{\pm 0.52}$	$\mathbf{4.68}_{\pm \mathbf{0.37}}$	$\mathbf{4.32}_{\pm \mathbf{0.53}}$	$\mathbf{4.86}_{\pm \mathbf{0.34}}$

Table 3: **Human evaluation** 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	4.03 / **1.83**	0.991	<0.001
Cardiology	93.3	3.73 / **1.67**	0.905	<0.001
General Surgery	100.0	3.80 / **1.57**	0.908	<0.001

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 archi-

tecture 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 +22.8/+25.3 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 reward-guided training; removing any one of these components reduces factuality below the full model (Table 7), indicating that each contributes to the observed gains. Full settings and analyses are provided in Table 4 and Appendices G, H, and L.

Table 4: Backbone generalizability of REMEMORIZE. Each backbone is evaluated in two configurations: without the memory module (backbone only) and with it (REMEMORIZE). Bold indicates the better value within each backbone–dataset pair. 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	**82.7**	51.2
	REMEMORIZE	MIMIC	**29.1**	**6.8**	**16.3**	79.3	**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	27.1	5.1	13.1	82.2	51.5
	REMEMORIZE	MIMIC	**31.1**	**7.1**	**18.8**	**83.9**	**74.3**
	Backbone only	SOAP	27.4	8.0	14.9	82.8	46.5
	REMEMORIZE	SOAP	**30.7**	**18.2**	**17.9**	**84.2**	**71.8**

6 CONCLUSION

We presented REMEMORIZE, a framework that regulates memory updates during autoregressive generation through gated recurrent memory and reward-guided memory-interface training. Across two clinical summarization benchmarks, it achieves higher factuality than the evaluated baselines. Ablations show lower factuality when components of the memory mechanism or reward-guided training are removed. Together, these results support explicit memory regulation as an approach to improving factual grounding in the evaluated clinical summarization tasks.

7 LIMITATIONS

We evaluate REMEMORIZE on two clinical summarization benchmarks; its generalizability to other domains and tasks remains unexplored. The recurrent memory architecture and reward-guided training objective impose no clinical priors and could, in principle, apply to other generation tasks requiring factual grounding. The reward function relies on NLI-based scoring, which may introduce bias in specialized clinical settings. Because sampled gate actions do not affect candidate generation, the auxiliary gate objective provides no informative candidate-reward signal for the gate in expectation. The reward-weight sensitivity analysis (Appendix K) uses one training run per configuration, so small differences between adjacent settings should be interpreted cautiously. We do not evaluate robustness under real-world deployment conditions. Residual hallucinations could contribute to clinical decision errors if summaries are used without human oversight. Our framework, however, is intended for research purposes and could be deployed in clinical settings only after further validation.

AI USE STATEMENT

AI-based tools (a large language model assistant) were used to support grammar and language editing of the manuscript, and to assist with implementing and debugging portions of the experimental code and providing feedback on aspects of the experimental methodology. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the accuracy and originality of this submission.

REFERENCES

Asad Aali, Dave Van Veen, Yamin Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, Jangwon Kim, and Akshay Chaudhari. MIMIC-IV-Ext-BHC: Labeled Clinical Notes Dataset for Hospital Course Summarization. *PhysioNet*, February 2025. doi: 10.13026/5gte-bv70. URL https://doi.org/10.13026/5gte-bv70. Version 1.2.0.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 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*, pp. 2291–2302, 2023.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *NPJ digital medicine*, 8(1):274, 2025.

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

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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)*, pp. 1724–1734, 2014.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 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*, pp. 1419–1436, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Peng Cui and Le Hu. 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*, pp. 5881–5891, 2021.

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

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

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

Ahmed Elhady, Khaled Elsayed, Eneko Agirre, and Mikel Artetxe. Improving factuality in clinical abstractive multi-document summarization by guided continued pre-training. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 755–761, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Nianlong Gu, Elliott Ash, and Richard H. R. Hahnloser. 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*, pp. 6507–6522, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 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*, pp. 3905–3920, 2022.

Ziwei Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 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)*, pp. 284–294, 2018.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

Younes Labrak et al. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

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

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

Chuyuan Li, Austin Xu, Shafiq Joty, and Giuseppe Carenini. Topic-guided reinforcement learning with LLMs for enhancing multi-document summarization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12395–12412, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.662. URL https://aclanthology.org/2025.findings-emnlp.662/.

Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. Clicare: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31554–31562, 2026.

Yinheng Li. 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*, pp. 641–647, 2023.

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

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

Shenling Liu, Yang Gao, ShaSha Li, PanCheng Wang, and Ting Wang. A hallucination detection and mitigation framework for faithful text summarization using llms. *Scientific Reports*, 16(1):1374, 2025.

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

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, pp. 9802–9822, 2023.

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

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

Xiaoyong Mei, Tingting Zuo, Da Chen, Guangyu Hu, Xiangyu Wen, Chao Duan, Mingyan Zhang, and Fudan Zheng. Beyond overlap metrics: Rewarding reasoning and preferences for faithful multi-role dialogue summarization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23189–23203, 2026.

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

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. 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)*, pp. 2653–2672, 2025.

Subash Neupane, Himanshu Tripathi, Shaswata Mitra, Sean Bozorgzad, Sudip Mittal, Shahram Rahimi, and Amin Amirlatifi. Clinicsum: Utilizing language models for generating clinical summaries from patient-doctor conversations. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 5050–5059. IEEE, 2024.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

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

Fahmida Liza Piya and Rahmatollah Beheshti. 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, pp. 29293–29294, 2025.

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

Khizar Qureshi, Geoffrey Martin, and Yifan Peng. Budget-aware routing for long clinical text. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42583–42598, 2026.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

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

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

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

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

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6252–6272, 2023.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 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)*, pp. 5858–5871, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.319. URL https://aclanthology.org/2024.acl-long.319/.

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

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

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 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)*, pp. 8371–8384, 2024.

Junjie Song, Yiwen Liu, Dapeng Li, Yin Sun, Shukun Fu, Siqi Chen, and Yuji Cao. Balancing rewards in text summarization: Multi-objective reinforcement learning via hypervolume optimization. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16777–16781. IEEE, 2026.

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

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric

Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

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

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639: 130193, 2025.

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

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

Nitin Yadav, Changsung Kang, Hongwei Shang, and Ming Sun. Generating query-relevant document summaries via reinforcement learning. *arXiv preprint arXiv:2508.08404*, 2025.

Lingrui Yang, Yuxing Zhou, Jun Qi, Xiantong Zhen, Li Sun, Shan Shi, Qinghua Su, and Xuedong Yang. Aligning large language models with radiologists by reinforcement learning from ai feedback for chest ct reports. *European journal of radiology*, 184:111984, 2025a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025b.

Cheng Yuan, Xinkai Rui, Yongqi Fan, Yawei Fan, Boyang Zhong, Jiacheng Wang, Weiyan Zhang, and Tong Ruan. LCDS: A logic-controlled discharge summary generation system supporting source attribution and expert review. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.),

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp. 284–294, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.28. URL https://aclanthology.org/2025.acl-demo.28/.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

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

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

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

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 continuous 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, group-relative advantages are computed following Shao et al. (2024); the gate parameters θ_π are updated via the clipped score-weighted auxiliary gate objective, and the memory parameters θ_M are updated separately via reward-weighted candidate learning.

Algorithm 1: REMEMORIZE Training Procedure

Input: Document d, group size G
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 $\mathrm{lm_head}(\tilde{h}_{t'})$;
    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 $\mathcal{L}_{\text{gate}}$ (Eq. 12);
Update θ_M via AdamW using $\mathcal{L}_{\text{mem}}$ (Eq. 8);

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 (Grattafiori et al., 2024) ($\mathcal{J}_{\text{L}}$), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) ($\mathcal{J}_{\text{M}}$), and Qwen2.5-7B-Instruct (Qwen Team et al., 2025) ($\mathcal{J}_{\text{Q}}$), 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:

$$\bar{s}_d(x) = \frac{1}{|\mathcal{J}^*(x)|} \sum_{j \in \mathcal{J}^*(x)} s_{j,d}(x). \tag{13}$$

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

$$\bar{S}_d = \frac{1}{N} \sum_{x=1}^{N} \bar{s}_d(x). \tag{14}$$

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 fourteen 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.

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`
LLaMA-3.1-8B	8B	`meta-llama/Llama-3.1-8B-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`

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 (Qwen Team et al., 2025), 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 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 of these general-purpose baseline models.

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 (Guo et al., 2025), 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. Thirty individuals participated voluntarily in the human evaluation study. The large majority of participants held advanced academic or clinical degrees (13 MBBS or equivalent; 11 MSc/MPH/MEng; 5 doctoral degrees), with one participant holding a bachelor's degree, yielding a final analytic sample of $n = 30$ respondents.

Procedure. Each respondent evaluated all three cases in a fixed presentation order, with cases separated by page breaks. For each case, an excerpt of the de-identified 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:

$$r \;=\; 1 \;-\; \frac{2W}{n(n+1)/2}, \tag{15}$$

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 $n = 30$ respondents. REMEMORIZE was identified as the less hallucinated summary in 90%, 93.3%, 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 ($r_{\mathrm{rb}} \geq 0.90$, $p \leq 0.001$). Pooling across all 90 respondent–case pairs, 94.4% of responses favored REMEMORIZE ($W = 135.5$, $p < 0.001$, $r_{\mathrm{rb}} = 0.931$).

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.

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
Memory dimension d_m	128
Key/value dimension (d_k, d_v)	(64, 64)
Gate bias init	-2.0
Reward weights $(\alpha, \beta, \gamma, \delta)$ – MIMIC-IV-Ext-BHC	(0.7, 0.5, 0.3, 1.2)
Reward weights $(\alpha, \beta, \gamma, \delta)$ – SOAP	(0.7, 0.9, 0.1, 0.4)
λ_{aux}	0.1

F.1 DATA EXTRACTION

All experiments are conducted using publicly available clinical summarization datasets. The MIMIC-IV-Ext-BHC benchmark (Aali et al., 2025) 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

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`. Values are reported as mean ± bootstrap standard deviation.

Variant	Rouge-1	Rouge-L	Factuality↑
Single-candidate variant	$19.5_{\pm 0.3}$	$9.5_{\pm 0.2}$	$50.8_{\pm 0.8}$
w/o reward-guided training	$19.2_{\pm 0.4}$	$9.4_{\pm 0.2}$	$51.0_{\pm 0.8}$
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	$27.1_{\pm 0.3}$	$13.1_{\pm 0.2}$	$51.5_{\pm 0.7}$
GRPO-direct (no memory)	$\mathbf{33.3}_{\pm 0.4}$	$\mathbf{19.3}_{\pm 0.2}$	$64.1_{\pm 0.7}$
Full model	$31.1_{\pm 0.3}$	$18.8_{\pm 0.2}$	$\mathbf{74.3}_{\pm 1.7}$

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 reward-guided training 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 reward-guided training contribute meaningfully to source-grounded generation.

Training variants. To evaluate reward-guided optimization, we modify its training components while preserving the memory architecture. The single-candidate variant uses $G = 1$ instead of a group of candidate summaries. The w/o reward-guided training variant removes reward-weighted candidate learning and the score-weighted auxiliary gate objective, retaining supervised summarization training. These variants assess the contribution of the combined reward-guided procedure; they do not isolate every reward-dependent component individually. Both obtain lower factuality scores than the full model. We also train a memory-free baseline, GRPO-direct, by applying group-relative policy optimization to backbone generation through LoRA under the same composite reward. GRPO-direct reaches 64.1 factuality, compared with 51.5 for the unmodified backbone and 74.3 for the full model. It achieves the highest ROUGE-1 and ROUGE-L scores in this table.

Backbone reference. The Backbone LLM row evaluates the unmodified `LLaMA-3.1-8B-Instruct` without memory augmentation or reward-guided training. It provides a reference for comparing the trained variants with the underlying pretrained model.

Inference-time memory interventions. We test two changes to the same trained ReMemorize checkpoint. The w/o Adaptive gate variant replaces the learned input-dependent gate with a fixed value of $\gamma_t = 0.5$ during inference. The w/o Phase-2 update variant retains the encoded source memory snapshot but stops memory updates during decoding. Both interventions reduce performance relative to the full model, indicating that its inference behavior depends on adaptive gating and continued memory updates. Because the checkpoint was trained with both mechanisms, these results do not establish how separately retrained models without those mechanisms would perform.

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 gated recurrent memory module, gating network, reward formulation, and reward-guided training procedure, are kept identical across experiments. This setting isolates the contribution of the proposed memory and gating 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 $\rightarrow$ 64.3) and +8.4 points on SOAP (61.4 $\rightarrow$ 69.8). On MIMIC-IV-Ext-BHC, ROUGE scores improve alongside factuality, while BERTScore F1 decreases slightly (82.7 $\rightarrow$ 79.3); on SOAP, R-L decreases slightly (25.7 $\rightarrow$ 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 +22.8 on MIMIC-IV-Ext-BHC and +25.3 on SOAP. On SOAP in particular, ROUGE scores increase substantially (R-1: 27.4 $\rightarrow$ 30.7; R-L: 14.9 $\rightarrow$ 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 NLI REWARD SCORING

The factuality and hallucination reward components (Eq. 4) are computed using a DeBERTa-v3-base cross-encoder fine-tuned for natural language inference (He et al., 2021), distinct from the MiniCheck model used for evaluation (Appendix J).

Prior NLI-based consistency methods such as SummaC segment source documents and aggregate inference scores across text units (Laban et al., 2022); our reward scorer instead uses overlapping source windows as described below.

Windowed scoring. The cross-encoder accepts at most 512 tokens per (premise, hypothesis) pair, while MIMIC-IV-Ext-BHC source documents average 2,267 tokens. To score the generated summary s^* against the complete source document d rather than a truncated prefix, d is partitioned into overlapping windows of $w = 450$ tokens with stride $r = 250$ tokens (a 200-token overlap between consecutive windows). Each window d_i is scored independently against s^*:

$$p_i^{\text{ent}} = P_{\text{NLI}}(\textit{entailment} \mid d_i,\, s^*), \qquad p_i^{\text{con}} = P_{\text{NLI}}(\textit{contradiction} \mid d_i,\, s^*). \tag{16}$$

The document-level reward components take the maximum over windows,

$$R_{\text{factual}} = \max_i\, p_i^{\text{ent}}, \qquad R_{\text{hallucination}} = \max_i\, p_i^{\text{con}}, \tag{17}$$

so a claim in s^* is credited as entailed if any window of d supports it, and flagged as contradicted if any window conflicts with it.

J 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 $(\text{document}, \text{claim})$ 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. Because MiniCheck accepts a limited context length per (document, claim) pair while MIMIC-IV-Ext-BHC source documents average 2,267 tokens, D is partitioned into overlapping windows of $w = 500$ tokens with stride $r = 250$ tokens (a 250-token overlap between consecutive windows), following the same windowing procedure as the NLI reward scorer (Appendix I). For each proposition c_k, MiniCheck computes a support probability against every window D_j, and the per-claim support probability is taken as the maximum over windows:

$$p_k = \max_j \Pr\big[\textsc{Supported} \mid D_j,\, c_k\big] \in [0, 1], \tag{18}$$

where $p_k > 0.5$ indicates that c_k is supported by some window of D. The per-summary factuality score is defined as the proportion of supported propositions:

$$\mathrm{MC}(\hat{y}, D) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\big[p_k > 0.5\big]. \tag{19}$$

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

$$\overline{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{MC}(\hat{y}^{(i)}, D^{(i)}) \in [0, 1]. \tag{20}$$

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.

Grounded factuality evaluation. MiniCheck assesses whether a generated claim is supported by a grounding document (Tang et al., 2024). We use it as a separate evaluation model from the DeBERTa-based NLI scorer used for training rewards. Both methods assess support from source evidence, but use different models and scoring procedures.

We use the `MiniCheck-Flan-T5-Large` checkpoint, which outperformed fact-checking systems of comparable size on the LLM-AggreFact benchmark (Tang et al., 2024). Inference is performed using a single forward pass per (window, claim) pair with greedy decoding. Bootstrap standard deviations (1000 resamples) are reported alongside the MiniCheck scores.

K Reward Weight Sensitivity Analysis

To assess the sensitivity of ReMemorize to the reward-weight hyperparameters, we conduct a coordinate-wise grid search over all four composite reward weights: α (factuality), δ (hallucination penalty), β (coherence), and γ (completeness). Reward configurations are not assumed to transfer across clinical domains, so all searches are performed independently on MIMIC-IV-Ext-BHC and SOAP.

The search procedure consists of four sequential sweeps per dataset, each varying one coefficient while the remaining three are held at their default values: $\alpha \in \{0.5, 0.7, 1.0, 1.3, 1.6\}$ with δ fixed at 0.8; $\delta \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ with α fixed at α^*; $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; and $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Each configuration is trained for one epoch on 300 samples, with model selection by MiniCheck factuality on the corresponding validation split. Figures 2 and 3 summarize validation factuality with bootstrap standard deviations across all four sweeps.

MIMIC-IV-Ext-BHC. Factuality peaks at $\alpha^* = 0.7$ (Fact $= 53.90 \pm 0.84$), above the default $\alpha = 1.0$ (Fact $= 51.74 \pm 0.84$), which is the weakest point in the sweep; the remaining values, including two grid-extension points at $\alpha = 0.9$ and $\alpha = 1.1$ (Fact $= 53.2$, 52.2), cluster within a narrow band (52.17–53.83). Fixing $\alpha = 0.7$, the δ sweep shows similarly narrow variation, peaking at $\delta^* = 1.2$ (Fact $= 53.34 \pm 0.83$) and dipping at $\delta = 1.0$ (Fact $= 51.80 \pm 0.84$; default $\delta = 0.8$: 52.79 ± 0.85), with no monotonic trend. For the auxiliary coefficients, factuality peaks at $\beta^* = 0.5$ (Fact $= 53.46 \pm 0.83$), above the default $\beta = 0.3$ (Fact $= 52.97 \pm 0.78$), with the weakest point at $\beta = 0.7$ (Fact $= 51.69 \pm 0.76$). For γ, factuality peaks at $\gamma^* = 0.3$ (Fact $= 52.70 \pm 0.75$); the remaining values cluster narrowly (51.78–52.56), with no monotonic trend.

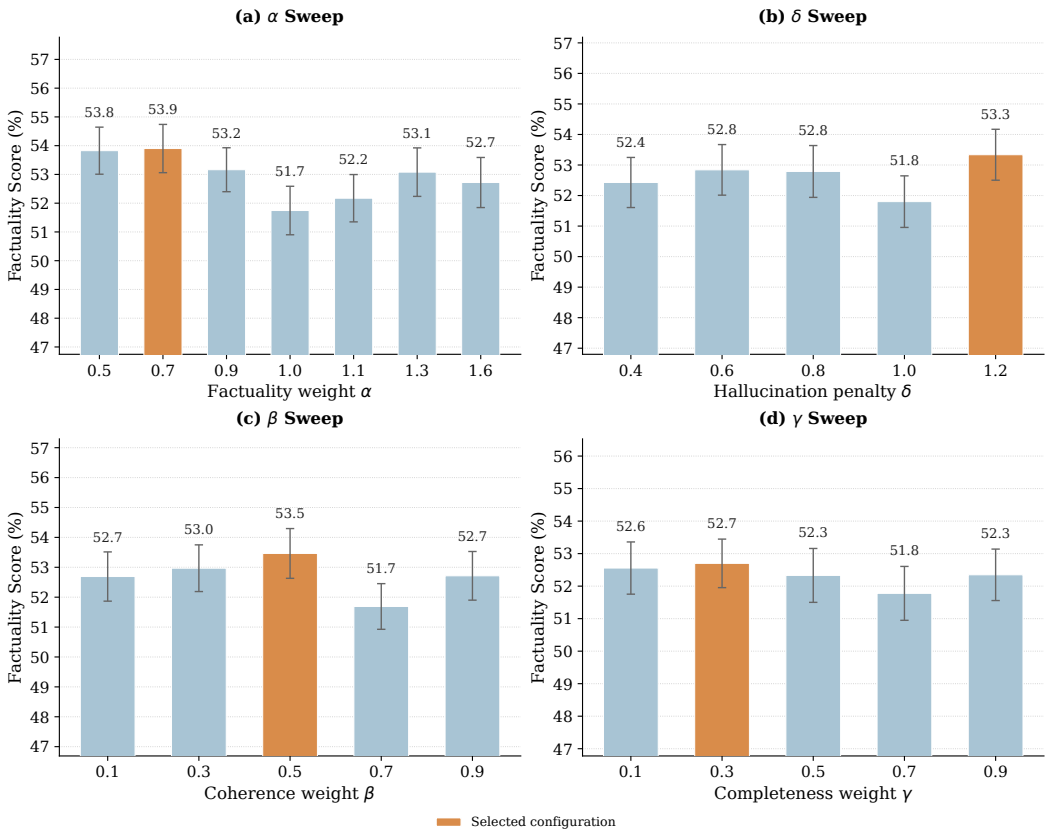

Figure 2: Coordinate-wise reward-weight sensitivity on the MIMIC-IV-Ext-BHC validation set. (a) Factuality across seven values of α with δ fixed at 0.8, including two grid-extension points beyond the standard range. (b) Factuality across five values of δ with α fixed at α^*. (c) Factuality across five values of β with α, δ fixed at their defaults. (d) Factuality across five values of γ, likewise. The selected configuration in each sweep (orange) also achieves the highest score. Error bars denote bootstrap standard deviation over 1,000 resampling iterations.

SOAP. $\alpha^* = 0.7$ achieves the strongest validation factuality (Fact $= 49.10 \pm 1.65$), matching the MIMIC-optimal value; other points cluster closely (Fact 47.97–49.02). For δ, $\delta^* = 0.4$ performs best (Fact $= 51.70 \pm 1.64$), effectively tied with $\delta = 0.8$ (Fact $= 51.60 \pm 1.59$); $\delta = 1.0$ falls in between (50.24 ± 1.58), while $\delta = 0.6$ and $\delta = 1.2$ score lowest (48.42, 48.21), with no monotonic trend. For the auxiliary coefficients, factuality peaks at $\beta^* = 0.9$ (Fact $= 50.68 \pm 1.64$), above the default $\beta = 0.3$ (Fact $= 48.85 \pm 1.68$); the extended grid points $\beta = 1.1$ and 1.3 score lower (49.80, 50.15). For γ, factuality peaks at $\gamma^* = 0.1$ (Fact $= 50.82 \pm 1.49$), coinciding with the default value; the remaining points cluster closely (48.51–49.29), with no monotonic trend.

L Memory Gate Activation Analysis

To better understand the behavior of the gate, we analyze the gate activations, encoded memory statistics, and auxiliary memory-consistency 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 memory-consistency 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.

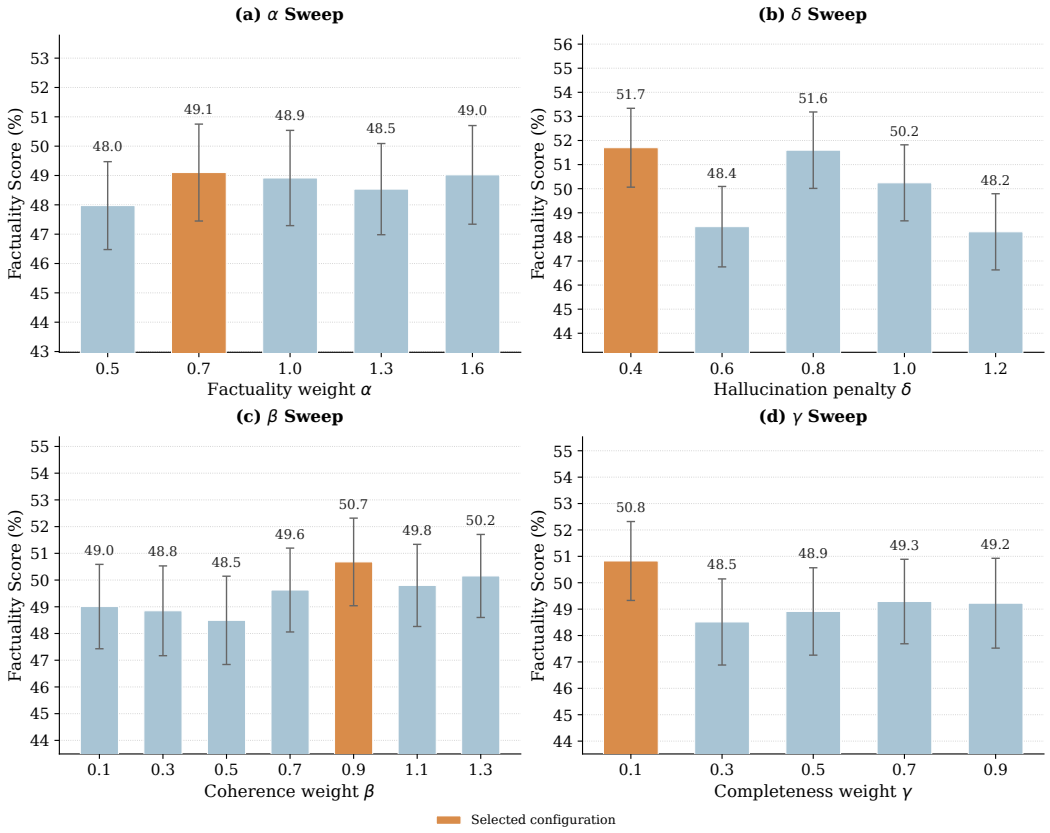

Figure 3: Coordinate-wise reward-weight sensitivity on the SOAP validation set. (a) Factuality across five values of α with δ fixed at 0.8. (b) Factuality across five values of δ with α fixed at $\alpha^* = 0.7$. (c) Factuality across seven values of β, including two grid-extension points beyond the standard range. (d) Factuality across five values of γ. The selected configuration in each sweep (orange) also achieves the highest score. Error bars denote bootstrap standard deviation over 1,000 resampling iterations.

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 memory-consistency 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
Single-candidate variant	0.098	5.43	0.271
w/o reward-guided training	0.091	5.37	0.271

Gate Activation Diagnostics. Across the three variants, mean gate activation is low ($\bar{\gamma} = 0.091$–0.098). Under the gated memory update in Eq. 1, low activation places greater weight on the previous memory state than on the candidate memory state. The similar means do not indicate a substantial change in average activation associated with gate training.

Impact of reward-guided optimization. The strongest contrast in Table 8 emerges between the full model and the variants trained without reward-conditioned optimization. Although the average gate activation remains nearly unchanged across the full model, the single-candidate variant,

and the w/o reward-guided training condition, the encoded memory statistics differ substantially. Removing reward-conditioned optimization 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 memory-consistency loss. When reward-conditioned optimization 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 memory-consistency loss (0.068) relative to the full model.

Although the reduced memory-consistency loss may initially appear favorable, it does not indicate improved memory quality: a fixed gate applies the same blending weight to every token, which can simplify the write-projection's target and lower $\mathcal{L}_{\text{aux}}$ without requiring better organization of source content. Replacing the input-dependent gate with a fixed gate substantially reduces factuality, despite the fixed-gate variant achieving a lower memory-consistency error. Because the fixed gate also changes the update weight to $\gamma_t = 0.5$, this comparison does not isolate the effect of input dependence or establish reward-learned gating.

M HALLUCINATION CASE STUDY

We examine a qualitative example from the SOAP (MTS-Dialog) test set in which a baseline model reverses an explicitly negative clinical finding — a clinically consequential hallucination pattern. MedAlpaca-7B and REMEMORIZE receive the same source conversation under the evaluation setting used in the main experiments.

Source Conversation. The source conversation documents an examination of the cervical lymph nodes:

> Doctor Shami, can you please take down the notes as I examine the patient?
> *Guest clinician:* Sure. Can you put the heading as cervical lymph nodes?
> *Guest clinician:* Okay. I am going to check for any swelling in your neck, okay? Can you look down? Okay. Alright, I do not feel anything here. Can you look up? Sure. No palpable nodes here. None of the lymph nodes are enlarged or have any swellings.
> *Guest clinician:* So, no lymphadenopathy?
> Yeah.

The clinically relevant evidence is an explicitly negative finding: *no palpable or enlarged cervical lymph nodes*.

MedAlpaca-7B transforms the absence of lymphadenopathy into a positive clinical history and recommends laboratory evaluation despite no such plan being discussed. This is not merely an omission: it reverses the clinical meaning of the source and introduces an unsupported intervention. In contrast, REMEMORIZE retains the documented negative finding and avoids adding a diagnosis or recommendation; although its output is more verbose than the reference summary, its central clinical claims remain aligned with the source evidence. This example illustrates, rather than establishes, the aggregate factuality gap reported in Table 1.

Table 9: Qualitative comparison of the reference summary, MedAlpaca-7B, and REMEMORIZE. The baseline reverses the negative finding and introduces an unsupported clinical recommendation, whereas REMEMORIZE preserves the finding expressed in the source.

Output	Generated summary	Relation to source
Reference	**Subjective:** Patient has concerns regarding cervical lymph nodes. No symptoms of swelling or lymphadenopathy mentioned. **Objective:** No swelling or palpable nodes in the neck. No enlarged lymph nodes or swellings observed. **Assessment:** No evidence of lymphadenopathy. **Plan:** No specific plan mentioned.	Consistent with the documented negative finding.
MedAlpaca-7B	**HPI:** History of lymphadenopathy. **RX:** Labs for evaluation of lymphadenopathy.	Reverses the negative finding and introduces an unsupported evaluation plan.
REMEMORIZE	**Chief complaint:** The patient's chief concern is not explicitly stated in the conversation, but it appears to be related to cervical lymph nodes. **Key findings:** No palpable swelling or enlargement of any lymph nodes was identified during the examination. **Assessment:** No lymphadenopathy was noted during the physical examination. **Plan:** No specific next steps were discussed.	Preserves the negative finding and introduces no unsupported diagnosis or recommendation.