

Autoregressive vs. Diffusion Language Models for Clinical Text Summarisation

Carlos A. Conde Giron

Utrecht University
Utrecht, Netherlands
c.a.condegiron@students.uu.nl

Zexuan Li

Utrecht University
Utrecht, Netherlands
z.li28@students.uu.nl

Robert A. Bagheri

Utrecht University
Utrecht, Netherlands
a.bagheri@uu.nl

Ruurd J. A. Kuiper

University Medical Center Utrecht
Utrecht, Netherlands
r.j.a.kuiper@umcutrecht.nl

Abstract

Autoregressive (AR) language models dominate clinical text summarisation, yet their sequential generation mechanism remains vulnerable to compounding errors. Diffusion language models theoretically mitigate this via iterative, sequence-level denoising, but work in the medical domain has so far centred on image-grounded tasks, leaving their efficacy on clinical text under-explored. This paper presents an empirical comparison of an AR model (Llama 3.1 8B) and a masked diffusion model (LLaDA 8B) on the MultiClinSum dataset. To constrain generation, we introduce Entity-Pinned Decoding (EPD), extracting and pinning critical clinical terms during denoising to reduce factual drift. We evaluate both paradigms using a three-tier framework comprising source-grounded factuality, reference-based similarity, and clinical acceptability rated by two LLM judges. Task-specific fine-tuning has a substantially larger effect on source-grounded faithfulness than the choice of paradigm. After fine-tuning, LLaDA and Llama achieve comparable factuality, with neither paradigm dominating across all source-grounded metrics. EPD improves source grounding but does not significantly increase entity retention, and it costs abstractive synthesis and inference speed. On this task, fine-tuned diffusion models are competitive with autoregressive models.

1 Introduction

The widespread adoption of electronic health records (EHRs) has inadvertently exacerbated physician administrative burden (Sinsky et al., 2016), increasing interest in the automation of clinical text summarisation. While Large Language Models offer a promising solution, their clinical application presents unique challenges. Unlike general-domain summarisation where surface-level

fluency is often sufficient, clinical NLP demands strict factual consistency. A single hallucinated medication or omitted diagnosis in a discharge summary can lead to severe clinical consequences.

Large Language Models (LLMs) based on the causal, left-to-right autoregressive (AR) paradigm currently serve as the backbone for most clinical summarisation systems (Van Veen et al., 2024). However, sequential generation can compound errors: once an AR model generates an inaccurate token, later tokens are conditioned on it and the error can propagate into a factual hallucination. Diffusion language models (DLMs) offer a promising alternative (Li et al., 2022). By treating generation as an iterative, sequence-level denoising process, DLMs refine the output globally, theoretically mitigating the left-to-right error cascade (Austin et al., 2021). Diffusion language models have begun to reach the biomedical domain, though mainly in image-grounded settings such as biomedical visual understanding (Dong et al., 2026) and medical visual question answering (Liu et al., 2026). Their performance on text-only, terminology-dense clinical documents remains under-explored.

Evaluating these paradigms presents a structural challenge. While source-grounded metrics like SummaC (Laban et al., 2022) and QAFactEval (Fabbri et al., 2022) verify output against the source document, the field remains reliant on reference-based evaluation. However, using open-source datasets with unverified ground truths is risky, as even expert-written summaries can contain factual errors. Because standard metrics reward models for matching flawed references and penalise them for generating factually correct deviations, reference-dependent evaluation can be unreliable in this setting.

To bridge these gaps, we present an empirical

comparison of 8B-parameter autoregressive and masked diffusion language models for clinical summarisation. We benchmark the AR baseline, Llama 3.1 (8B) (Dubey et al., 2024), against an open-source masked diffusion model, LLaDA (8B) (Nie et al., 2025). We also introduce Entity-Pinned Decoding (EPD), an inference-time mechanism that pins critical clinical entities during denoising to reduce factual drift. Our evaluation uses three tiers ordered by clinical risk: source-grounded factuality, reference-based similarity as a secondary diagnostic, and clinical acceptability rated by two LLM judges with a rubric adapted from the PDSQI-9 (Croxford et al., 2025).

Our research questions are:

- **RQ1:** How do autoregressive and diffusion models compare under zero-shot and fine-tuned conditions?
- **RQ2:** Does Entity-Pinned Decoding measurably reduce entity-level factual drift in diffusion models?
- **RQ3:** How do these paradigms perform across the dimensions of factual safety and clinical utility, and what does a multi-tier evaluation reveal that single-metric benchmarking misses?

To address these questions, we make three contributions: (1) an empirical comparison of autoregressive and diffusion language models for clinical summarisation at the 8B scale, (2) Entity-Pinned Decoding, a logit-level guidance mechanism for entity preservation in masked diffusion, and (3) a three-tier evaluation combining source-grounded factuality, reference-based similarity, and clinical acceptability.

2 Related Work

Clinical text summarisation has moved from extractive methods and small sequence-to-sequence models (Zhang et al., 2018) to causal LLMs (Van Veen et al., 2024; Dubey et al., 2024; Jiang et al., 2023). These models are fluent, but surveys warn against their use in safety-critical settings because of factual inconsistency (Ji et al., 2023). Left-to-right generation suffers from exposure bias and compounding errors (Huang et al., 2023; Zhang et al., 2023), and in clinical text such errors can surface as hallucinated conditions or omitted symptoms

(Omiye et al., 2024; Pal et al., 2023). Decoding-time interventions have been proposed (Yang and Klein, 2021; Chuang et al., 2024), but the vulnerability rooted in token-by-token causal decoding remains, which motivates alternative generative paradigms.

Diffusion language models generate by iteratively denoising a full sequence rather than one token at a time (Li et al., 2022; Austin et al., 2021), which allows bidirectional context and global revision. Masked diffusion has since been scaled to the 8B regime: LLaDA (Nie et al., 2025) approaches the instruction-following ability of comparable AR models, and LAD (Kuiper et al., 2025) obtains diffusion-style generation by adapting a pretrained AR model with LoRA adapters. In the biomedical domain, diffusion language models have so far been applied mainly in image-grounded settings, such as biomedical image understanding (Dong et al., 2026) and medical visual question answering (Liu et al., 2026), both of which produce short answers conditioned on an image. We are not aware of prior work applying masked diffusion language models to long, terminology-dense clinical text.

Lexically constrained decoding guarantees the inclusion of specified phrases by extending autoregressive beam search (Hokamp and Liu, 2017; Post and Vilar, 2018). These methods restructure the search over the next-token distribution and have no counterpart in the parallel remasking step of masked diffusion, which our Entity-Pinned Decoding targets (Section 3.5). For entity identification we use the standard biomedical pipeline of scispaCy NER trained on BC5CDR (Neumann et al., 2019; Li et al., 2016) with UMLS linking (Bodenreider, 2004).

Lexical overlap metrics such as ROUGE and BLEU (Lin, 2004; Papineni et al., 2002) are known to be inadequate for clinical text, since they can assign high scores to factually wrong summaries (Maynez et al., 2020; Adams et al., 2023; Goyal et al., 2022). Source-grounded metrics address this: SummaC (Laban et al., 2022) tests sentence-level entailment against the source, and QAFactEval (Fabbri et al., 2022) checks factual consistency through question answering, improving over earlier QA-based metrics by 14% on the SummaC benchmark, with QA-based and NLI-based signals combining for a further gain. In clinical text specifically, standard automatic metrics correlate weakly with expert judgement (Tang et al., 2023). LLM-as-a-judge protocols (Zheng et al., 2023) offer a

scalable complement: an LLM jury reached almost perfect agreement ($\kappa = 0.81$) with a 7-physician majority vote on the presence of clinician-defined key facts (Grolleau et al., 2026), and a panel of judges from different model families aligns better with human judgement than a single large judge (Verga et al., 2024). We combine these three signals in a single evaluation.

3 Methodology

3.1 Dataset

We use MultiClinSum (Rodríguez-Ortega et al., 2025), a publicly available multilingual clinical summarisation dataset. We train on the combined gold-standard and large-scale English splits (26,494 pairs: 592 GS, 25,902 LS) and evaluate on a subset of the English test split (175 pairs). Each instance consists of a full clinical case report and a human-written reference summary.

In the GS split, source notes average 562 words (mean 3,786 characters) and reference summaries average 104 words, giving a compression ratio of approximately 5:1. The LS split has comparable length characteristics to the GS split. Tokenised using the Llama 3.1 tokeniser (meta-llama/Llama-3.1-8B), 80.2% of GS source documents exceed 512 tokens and 25.5% exceed 1,024 tokens.

3.2 Reference Summary Reliability

We audit the 592 GS training pairs for entity-level consistency between each source note and its reference summary, using the scispaCy entity extraction pipeline described in Section 3.5. Of the 592 pairs, 58 (9.80%) are flagged by the pipeline as containing at least one inconsistency. The types are medications in the reference that are absent from the source (43 pairs, 7.26%), numeric clinical value changes under the same unit (10 pairs, 1.69%), patient age mismatches exceeding two years (5 pairs, 0.84%), and gender contradictions (3 pairs, 0.51%). The subcategory counts exceed 58 because three pairs contain more than one inconsistency type. Table 5 in Appendix E shows three examples.

Because close to one in ten reference summaries disagrees with its source note on a clinical entity, we evaluate factual consistency against the source document directly rather than against the reference (Section 3.6), and we pin source entities during generation (Section 3.5).

3.3 Models

Llama 3.1 8B (Dubey et al., 2024) serves as the autoregressive baseline, loaded from meta-llama/Llama-3.1-8B. It generates text left-to-right using standard causal language modelling, predicting each token conditioned on all previous tokens. The fine-tuned condition applies the LoRA setup of Section 3.4 to this base checkpoint.

LLaDA 8B (Nie et al., 2025) is a masked diffusion language model loaded from GSAI-ML/LLaDA-8B-Base, trained from scratch at 8B scale. Rather than sequential generation, it begins from a fully masked sequence and iteratively denoises it: at each step, tokens with the highest predicted confidence are unmasked and fixed, while low-confidence tokens are remasked for further refinement. LLaDA has a Transformer architecture similar in size and depth to Llama 3.1 8B, but without causal masking. Matching scale removes one confound, although the two models still differ in pretraining data and tokeniser.

We evaluate each model under two conditions, zero-shot and fine-tuned, using different checkpoints for each. Zero-shot inference uses the instruction-tuned (Instruct) checkpoints, since the base checkpoints without instruction tuning did not produce usable zero-shot output, while the fine-tuned condition applies the LoRA adapter of Section 3.4 to the base checkpoints. Full checkpoint and training details are given in Appendix B.

3.4 Fine-Tuning Setup

Both models are fine-tuned on the combined GS and LS training splits of MultiClinSum, comprising 26,494 English clinical note and summary pairs (592 GS, 25,902 LS). The LoRA configuration, optimiser, schedule, batch size, and number of epochs are kept identical across the two models, so that performance differences reflect the generative paradigm rather than the training procedure.

We apply LoRA (Hu et al., 2022) with rank $r = 8$, $\alpha = 32$, dropout 0.1, targeting the query, key, value, and output projection matrices (q_proj , k_proj , v_proj , o_proj). Only this adapter is trained. The 8B base and the language modelling head are frozen. Training runs for 3 epochs with a batch size of 1 and gradient accumulation over 8 steps, an effective batch size of 8, using AdamW with learning rate 2×10^{-4} , weight decay 0.01, 20 linear warmup steps, and gradient clipping at norm 1.0. These follow commonly used settings

for LoRA fine-tuning at this scale and were not independently tuned. Training pairs whose prompt exceeded 2,048 tokens were excluded, affecting 187 Llama and 173 LLaDA pairs out of 26,494 (0.71% and 0.65%). The small difference follows from tokeniser vocabulary and prompt template.

Llama 3.1 8B is loaded with `AutoModelForCausalLM` in `bfloat16`, with the padding token set to the EOS token as Llama 3.1 has no dedicated padding token. It is trained with the standard causal language modelling loss over the summary tokens only. LLaDA 8B is loaded with `AutoModel` in `bfloat16`, with the mask token ID set to 126336. It follows the masked diffusion objective of Nie et al. (2025): for each example a masking ratio $t \sim \mathcal{U}[10^{-3}, 1)$ is applied to the summary tokens and the model reconstructs the original tokens given the full source note.

Software dependencies for both models are listed in Appendix A.

3.5 Entity-Pinned Decoding

Standard lexically constrained decoding (Hokamp and Liu, 2017) forces a set of words to appear in the output by extending autoregressive beam search, which has no counterpart in masked diffusion’s parallel decoding. We instead add a bonus to the logit of each entity’s first token across all output positions at each denoising step, following logit-level guidance for controllable generation (Yang and Klein, 2021). The bonus is largest when the sequence is fully masked and decays to zero by the final steps.

Disease and medication spans are identified by the `en_ner_bc5cdr_md` scispaCy model (Neumann et al., 2019), trained on the BC5CDR corpus (Li et al., 2016). Medication spans are filtered by UMLS Semantic Type (Bodenreider, 2004) to keep only pharmacological substances and clinical drugs, removing laboratory analytes and endogenous chemicals. Entities longer than ten tokens are discarded, which removes occasional NER errors where a full clause is tagged as a single disease. Ages, vital signs, dosages, and gender references are extracted via regex. Linked entities are resolved to UMLS canonical names (Bodenreider, 2004) so that surface variants such as *paracetamol* and *acetaminophen* refer to the same concept.

Source documents in the GS split contain a mean of 15.7 diseases, 5.3 medications, and 7.7 clinical numbers per document, but reference summaries retain only 23.6%, 11.2%, and 30.5% of these re-

spectively. Gender is retained in 85.7% of reference summaries and is always pinned. Pinning everything else would fill the generation budget with unimportant entities, so medications and diseases are ranked and a subset is selected. Medications are ranked by normalised mention frequency in the source document, with a $2\times$ multiplier when the entity’s last mention falls in the final quarter, since late mentions in problem-oriented notes tend to sit in Assessment and Plan sections (Weed, 1968). Diseases are additionally weighted by normalised inverse document frequency over the training corpus (Sparck Jones, 1972), which down-weights frequent comorbidities such as *pain* relative to rare, specific diagnoses, with the IDF table fitted once using Laplace smoothing (Manning et al., 2008).

Entities are selected in priority order: age, gender, top-scored medications, top-scored diseases, and remaining numbers, subject to a token budget of 30% of the target generation length (38 tokens at $L = 128$). Per-category caps ($K_{\text{med}} = 4$, $K_{\text{dis}} = 8$) are set above the number of medications and diseases that reference summaries typically retain. For multi-token entities, only the first token receives the bonus. In practice the remaining tokens tend to follow once the first is placed with high confidence. The full set of Entity-Pinned Decoding hyperparameters, including the logit bonus λ_{max} , is listed in Table 4 (Appendix B), and the implementation is available at <https://github.com/bagheria/DiffuClinic>.

3.6 Evaluation Methodology

Evaluation Framework: Surface-level n-gram matching is insufficient for the medical domain, where factual consistency is critical (Maynez et al., 2020), so we structure evaluation across three tiers ordered by clinical risk:

- **Tier 1: Source-Grounded Factuality.** We evaluate factual consistency at multiple granularities. For contradiction detection, we employ SummaC (Laban et al., 2022), and for question-based factual consistency, we use a lightweight QA metric that follows the design of QAFactEval (Fabbri et al., 2022) but differs in components and scoring (Appendix C). To complement these with entity-level analysis, we use MedNER-F1, applying an established clinical NER pipeline (Neumann et al., 2019) to quantify the exact preservation of critical medical entities.

- **Tier 2: Linguistic Similarity.** To measure surface-level fluency and alignment with reference summaries, we report standard token-matching metrics including ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). Furthermore, we incorporate BERTScore-F1 (Zhang et al., 2020) to capture broader semantic similarity beyond exact lexical overlap.
- **Tier 3: Clinical Acceptability.** Recognising the limitations of rigid programmatic metrics, we deploy an LLM-as-a-Judge evaluation to assess clinical utility. Motivated by recent validations of LLM juries for clinical text evaluation (Grolleau et al., 2026), we assess summaries using a rubric adapted from the PDSQI-9 (Provider Documentation Summarization Quality Instrument), reporting four factors: Accuracy, Completeness, Organisation, and Synthesis (Croxford et al., 2025). The full rubric and scoring anchors are given in Appendix D. We employ two LLM judges, o3-mini and DeepSeek-R1. Using judges from two model families provides a check against dependence on a single evaluator (Verga et al., 2024). All 175 test samples across all model configurations are evaluated by both judges, yielding 1,750 judgements in total.

Exact model versions and configurations for all metrics are provided in Appendix C.

Sample Size and Statistical Testing: The full MultiClinSum English test set contains 3,396 documents, and the NLI and QA-based metrics are too expensive to run on all of them. We therefore size the evaluation set with a pilot-based power analysis (Card et al., 2020; Browne, 1995; Sim and Lewis, 2012): a 50-sample pilot gave a largest standard deviation among the Tier 1 metrics of $\sigma = 0.269$ (SummaC), from which $n = 175$ yields a 95% confidence interval half-width of ± 0.04 and over 80% power to detect an effect of 0.075 (Appendix F). All main results use this fixed subset. Because the metrics are not normally distributed, we compute 95% confidence intervals with 1,000 bootstrap resamples and compare models with paired bootstrap tests (Deutsch et al., 2021; Dror et al., 2018).

4 Results

4.1 Experimental Setup

All models are evaluated on the 175-document test subset determined in Section 3.6, sized for a SummaC confidence interval half-width of ± 0.04 . Differences Δ are LLaDA minus Llama unless stated otherwise. Tier 1 and Tier 3 Overall scores are reported as the mean $\pm$ the 95% bootstrap margin of error ($k = 1,000$, random seed 42, see Table 1), Tier 2 metrics and the per-factor PDSQI-9 scores are reported as point estimates. Pairwise comparisons use paired bootstrap tests at $\alpha = 0.05$. All inference runs on a single NVIDIA A10 GPU (24 GB).

4.2 Main Results

4.2.1 Factuality

Table 1 reports the Tier 1 factuality scores. In the zero-shot setting, both LLaDA and Llama 3.1 register negative SummaC scores (-0.405 and -0.339), and their QAFactEval scores fall between 0.42 and 0.44, a difference that is not significant ($\Delta = -0.021$).

After clinical LoRA fine-tuning, SummaC becomes positive for both models (LLaDA $+0.039$, Llama $+0.078$), and the fine-tuned versus zero-shot difference is significant for each model under a paired bootstrap test ($p < 0.001$). QAFactEval reaches 0.608 for LLaDA LoRA and 0.561 for Llama LoRA ($\Delta = +0.047$ in favour of LLaDA), and MedNER-F1 is 0.380 for LLaDA LoRA and 0.386 for Llama LoRA, a difference that is not significant.

Entity-Pinned Decoding, applied to LLaDA LoRA, raises SummaC to $+0.150$ ($\Delta = +0.11$ over LLaDA LoRA, $p < 0.01$) and QAFactEval to 0.664 ($\Delta = +0.056$, $p < 0.05$). MedNER-F1 increases from 0.380 to 0.398, a change that is not significant. This is the highest MedNER-F1 of any configuration, and it is level with the 0.386 of Llama 3.1 LoRA.

4.2.2 Linguistic Similarity

Figure 1 reports the Tier 2 metrics. In the zero-shot setting, LLaDA reaches the highest BERTScore-F1 of any configuration (0.205) and leads Llama 3.1 on ROUGE-L (0.275 vs. 0.250, $\Delta = +0.025$, $p < 0.05$).

After clinical LoRA fine-tuning, Llama 3.1 leads on ROUGE-L (0.279), BLEU-4 (0.095), and METEOR (0.264), with the ROUGE-L difference sig-

Metric	Zero-Shot		LoRA Fine-tuned		EPD
	Llama-3.1	LLaDA	Llama-3.1	LLaDA	LLaDA EPD
<i>Tier 1</i>					
SummaC	$-0.339 \pm .039$	$-0.405 \pm .035$	$0.078 \pm .036$	$0.039 \pm .035$	$0.150 \pm .029$
QAFactEval	$0.439 \pm .026$	$0.418 \pm .026$	$0.561 \pm .032$	$0.608 \pm .033$	$0.664 \pm .033$
MedNER-F1	$0.381 \pm .030$	$0.357 \pm .024$	$0.386 \pm .034$	$0.380 \pm .033$	$0.398 \pm .035$
<i>Tier 3</i>					
Overall (o3)	$3.051 \pm .058$	$2.512 \pm .110$	$2.547 \pm .057$	$2.151 \pm .091$	$2.420 \pm .054$
Overall (R1)	$2.854 \pm .086$	$2.099 \pm .106$	$2.489 \pm .072$	$1.970 \pm .098$	$2.323 \pm .070$

Table 1: Main evaluation results across Tier 1 and Tier 3 metrics. Scores are presented as the mean value $\pm$ 95% bootstrap margin of error.

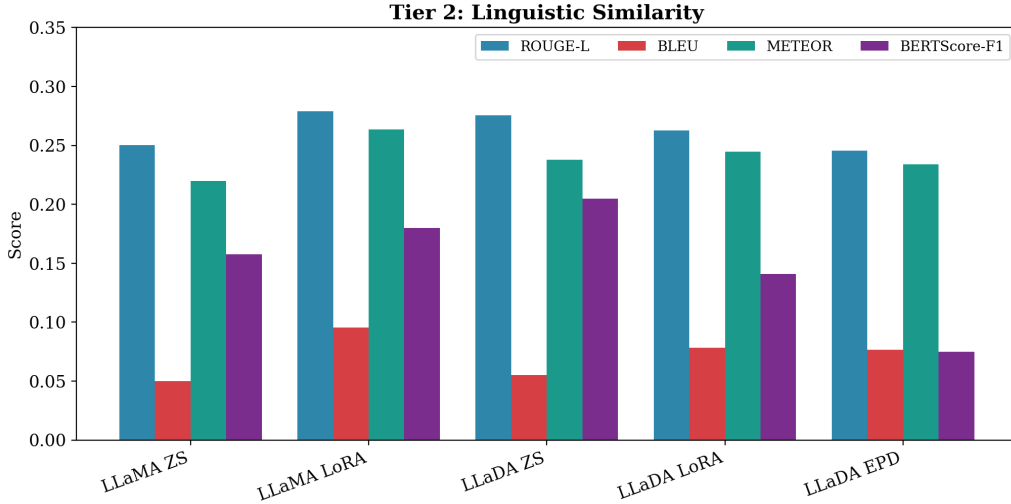


Figure 1: Tier 2 linguistic similarity scores across all model configurations. BLEU is BLEU-4.

nificant ($\Delta = -0.017$, $p < 0.05$). BERTScore-F1 moves in opposite directions for the two models, falling from 0.205 to 0.141 for LLaDA and rising from 0.158 to 0.180 for Llama 3.1.

Adding Entity-Pinned Decoding to LLaDA LoRA lowers ROUGE-L to 0.246, METEOR to 0.234, and BERTScore-F1 to 0.075, the lowest of any configuration. One of the 175 EPD generations is empty and receives a BERTScore-F1 of -4.93 . Excluding that single document, the mean is 0.104. We report the full-sample mean, since a decoding method that can produce an empty output is itself a failure mode worth counting. BLEU-4 does not exceed 0.10 in any configuration, with a maximum of 0.095 for Llama 3.1 LoRA.

4.2.3 Clinical Acceptability

Table 2 reports PDSQI-9 scores from both judges. The two judges agree on the ranking except for one adjacent pair, and o3-mini assigns systematically higher absolute scores than DeepSeek-R1.

In the zero-shot setting, Llama 3.1 receives the highest Overall score from both judges (3.05 from

o3-mini, 2.85 from DeepSeek-R1), driven mainly by Organisation (4.70 vs. 3.59 for LLaDA under o3-mini). Fine-tuning lowers the Overall score for both models (Llama 3.1 3.05 to 2.55, LLaDA 2.51 to 2.15 under o3-mini), with the largest drops in Completeness and Synthesis. Applying EPD to LLaDA LoRA raises its Overall score (2.15 to 2.42), with its largest gain in Organisation (3.36 to 4.00) and a further decline in Synthesis (0.74 to 0.60). Per-factor scores are in Table 2. Synthesis does not exceed 2.2 for any configuration.

4.3 Bootstrap Validation and Pairwise Comparisons

The post-hoc bootstrap half-width of the SummaC confidence interval is at most ± 0.039 , which is within the ± 0.04 target specified by the power analysis in Section 3.6, confirming that $n = 175$ is sufficient.

Paired bootstrap comparisons show that the fine-tuning effect is larger than the paradigm difference. Fine-tuning gains on SummaC ($+0.42$ to $+0.44$, $p < 0.001$) and QAFactEval ($+0.12$ to $+0.19$, $p <$

Table 2: Clinical acceptability scores by LLM judge on the rubric adapted from the PDSQI-9. Scale 1–5, higher is better.

Model	Accuracy	Completeness	Organisation	Synthesis	Overall
<i>o3-mini</i>					
Llama ZS	2.82	2.41	4.70	1.45	3.05
Llama LoRA	2.77	1.59	4.23	0.74	2.55
LLaDA ZS	2.03	2.55	3.59	1.33	2.51
LLaDA LoRA	2.37	1.53	3.36	0.74	2.15
LLaDA EPD	2.84	1.45	4.00	0.60	2.42
<i>DeepSeek-R1</i>					
Llama ZS	2.55	1.94	4.21	2.04	2.85
Llama LoRA	2.57	1.55	3.70	1.53	2.49
LLaDA ZS	1.45	2.06	2.90	1.59	2.10
LLaDA LoRA	1.86	1.52	2.75	1.37	1.97
LLaDA EPD	2.55	1.46	3.46	1.25	2.32

0.001) exceed the between-paradigm differences in every condition. On SummaC, the fine-tuning effect is six to eleven times the paradigm gap. The QAFactEval paradigm gap is small in both settings ($\Delta = -0.021$ zero-shot, $\Delta = +0.047$ fine-tuned), and MedNER-F1 shows no significant paradigm difference under any condition. SummaC shows a small autoregressive (AR) advantage ($\Delta = -0.039$ to -0.066 , $p < 0.05$). On ROUGE-L, the direction of improvement depends on training, with LLaDA leading zero-shot ($+0.025$, $p < 0.05$) and Llama 3.1 leading after fine-tuning (-0.017 , $p < 0.05$).

Because we compute many metrics and run many pairwise tests, we report p -values without correction for multiple comparisons, so individual comparisons should be read as descriptive rather than confirmatory.

5 Discussion

5.1 Lexical Overlap and Source Faithfulness Diverge

Lexical overlap and source-grounded faithfulness can point in opposite directions, so neither can stand alone in clinical evaluation. In the zero-shot setting, LLaDA scores higher than Llama 3.1 on ROUGE-L and BERTScore-F1 but lower on SummaC, meaning its summaries contradict the source more often than Llama’s. A reader who looked only at ROUGE or BERTScore would rank LLaDA first even though it is the less faithful of the two. Without task-specific training, the global, bidirectional denoising of the diffusion model appears to admit a wider range of valid surface expressions than left-to-right decoding, which is consis-

tent with its higher zero-shot BERTScore. After fine-tuning the disagreement takes a different form, with Llama 3.1 LoRA leading on ROUGE-L and BLEU-4 while LLaDA LoRA scores higher on QAFactEval, and the two source-grounded metrics disagree with each other, since Llama leads on SummaC while LLaDA leads on QAFactEval. This is a reason to report more than one faithfulness metric rather than relying on a single number.

N-gram metrics measure similarity to one reference, but a clinically useful summary must be consistent with the source note. Our qualitative reading (Appendix E) shows the fine-tuned models often copy source phrases, which raises ROUGE and BLEU while sometimes carrying text out of context, and the sub-0.10 BLEU-4 ceiling is likely a metric artefact, since clinical content can be phrased in many equivalent ways. We therefore treat source-grounded faithfulness as a minimum requirement for any improvement claim and do not rely on ROUGE or BLEU alone.

5.2 Reference Noise Motivates Source-Grounded Evaluation

Our audit of the MultiClinSum gold-standard training partition (Section 3.2) found that 58 of 592 reference summaries, 9.80%, disagree with their source note on at least one clinical entity, including gender, age, and numeric values. This affects evaluation design directly. If close to one in ten references is itself wrong on an entity, a reference-based metric (ROUGE, BLEU, METEOR, BERTScore) will penalise a model for being right when the reference is wrong. A model that writes *29-year-old female* where the reference wrongly says *29-year-old male* loses ROUGE-1 for the correct out-

put. Source-grounded metrics avoid this by scoring against the source note rather than the reference. We read this as a reason to report source-grounded metrics as the primary endpoints in clinical summarisation and to keep reference-based metrics as secondary checks, rather than as a reason to discard reference metrics entirely.

5.3 Trade-offs Between Diffusion and Autoregressive Models

Fine-tuning has a substantially larger effect on faithfulness than the choice of paradigm (Section 4.3). After fine-tuning the two models are close, with Llama leading on SummaC, LLaDA on QAFactEval, and the two level on entity retention, so where faithfulness is the main concern the choice between them can rest on other grounds.

Entity-Pinned Decoding adds a real inference cost. On the A10 used for our runs, LLaDA LoRA takes 5.97 seconds per sample with the dLLM-Cache feature cache and 18.4 seconds with EPD, which ran without the cache. The ratio of roughly $3.1\times$ is therefore an upper bound on the cost of the logit bonus itself, which we did not isolate.

The LLM-judge scores favour the zero-shot autoregressive model, which has the highest Overall PDSQI-9 score from both judges, driven mainly by Organisation. Fine-tuning lowers the judge scores for both paradigms, with the largest drops in Completeness and Synthesis, the opposite direction to the faithfulness metrics. The pattern is consistent with clinical LoRA training pushing the models towards extractive, source-copying behaviour, which raises measured faithfulness and n-gram overlap but reads as less complete and less synthesised. Entity-Pinned Decoding recovers some of the judge-rated organisation for the diffusion model but lowers Synthesis further, consistent with pinned entities acting as structural anchors while narrowing the vocabulary available for abstraction. Synthesis stays low across every configuration, which suggests a limit that is not specific to one architecture or decoding choice, and training a model that is both faithful to the source and well structured for a clinician remains open.

5.4 Answering the Research Questions

RQ1. The comparison depends on training. Zero-shot, the paradigms show complementary strengths, diffusion giving broader semantic coverage and the AR model better-structured narratives. After clinical LoRA fine-tuning the two are not significantly

different on entity retention, and the AR model keeps only a small entailment advantage that is minor next to the fine-tuning effect.

RQ2. The benefit of EPD is narrow. It improves sentence-level grounding and judge-rated organisation, but it does not significantly increase entity retention, a level the AR model reaches with no entity-aware mechanism, and it costs semantic richness, abstractive synthesis, and roughly three times the latency. EPD is a targeted tool for when source grounding is the priority, not a general-purpose improvement.

RQ3. No single tier captures the observed differences. Tier 1 alone would call the fine-tuned models close and miss the judges’ preference for zero-shot structure. Tier 2 alone ranks by surface overlap, which does not track faithfulness: the zero-shot Tier 2 leader (LLaDA) is the less faithful model on SummaC, and ROUGE gives no signal separating that case from the fine-tuned one, where the ROUGE leader is also faithful. Tier 3 alone would follow the judges towards a less faithful model. Together they make the trade-offs visible.

6 Conclusion

We compared autoregressive and masked diffusion language models for clinical text summarisation under a source-grounded, multi-metric evaluation. Task-specific fine-tuning has a substantially larger effect on faithfulness than the choice of paradigm, and after fine-tuning the two models are close with no single consistent winner. Entity-Pinned Decoding gives the highest source-grounded scores among the evaluated configurations, at the cost of lower reference-based scores, lower Synthesis, and roughly three times the inference latency. Because n-gram metrics and source-grounded faithfulness disagree at both training stages, and close to one in ten reference summaries contradicts its source note, we treat source-grounded faithfulness as the primary endpoint and reference-based metrics as secondary checks.

Limitations

Our main evaluation uses 175 test samples, about 5.2% of the MultiClinSum English test set. This subset was sized by a power analysis for the primary SummaC comparison and gives a 95% confidence interval half-width near ± 0.04 , but rare entities and unusual cases may be under-represented, and the smaller between-model differences such as

MedNER-F1 would benefit from a larger sample.

We report p -values without correction for multiple comparisons across a large number of metrics and pairwise tests, so the familywise error rate is inflated and some individual comparisons may reach significance by chance (Section 4.3). The power analysis in Appendix F sizes the sample for a single primary SummaC comparison rather than for the full family of tests. We therefore read the pairwise results as descriptive, and a confirmatory study would need a pre-specified primary endpoint with a correction such as Bonferroni.

Our clinical acceptability scores come from two LLM judges rather than clinicians. We use two models with different training to reduce single-judge bias, but known judge effects such as position, verbosity, and self-preference are not controlled, and the higher overall scores from o3-mini may reflect a scoring bias rather than real quality differences. Direct evaluation by clinicians would be the stronger test, which we did not run.

All experiments use MultiClinSum, which covers English case reports. The findings may not carry over to other clinical text types such as radiology reports or discharge summaries, and cross-dataset evaluation would be needed to test this.

Our experiments are limited to the 8B scale. Whether the same pattern, with fine-tuning dominating the paradigm gap on faithfulness, holds at larger scales is left open.

The zero-shot and fine-tuned conditions use different checkpoints (Section 3.3). Zero-shot runs use the Instruct checkpoints and fine-tuned runs use the base checkpoints with LoRA, so what we call the fine-tuning effect is the difference between these two stages, not the effect of adding LoRA to a fixed model. Part of it, including the drop in judge-rated Completeness and Synthesis, may reflect the loss of general instruction tuning rather than task adaptation.

Generation was capped at 128 tokens for every system, while 53% of the 175 reference summaries are longer than 128 Llama tokens, so every system is truncated relative to the references and the Llama runs reached the cap on nearly every document. Part of the drop in judge-rated Completeness after fine-tuning may reflect this cap. The non-EPD LLaDA runs used the dLLM-Cache feature cache and the EPD run did not, so the reported latency ratio is an upper bound on the cost of EPD itself. Each model was fine-tuned once without a fixed random seed, so the confidence intervals reflect

variation over test documents, not over training runs.

Entity-Pinned Decoding was tested in a single pipeline, LLaDA with fine-tuning, with EPD turned on and off (Section 3.5). This isolates the EPD effect for that model, holding fine-tuning fixed: it raises SummaC by +0.11 and QAFactEval by +0.06, and raises MedNER-F1 from 0.380 to 0.398. We did not test EPD with the autoregressive model or without fine-tuning, so its interaction with those settings is unknown. The MedNER-F1 gain is small, and the EPD hyperparameters were not tuned on a held-out validation set: the per-category caps follow the reference summary statistics (Section 3.5), the logit bonus was set from a small number of calibration runs, and the annealing schedule was left at its default. A systematic sweep over these parameters would show how sensitive the EPD gains are to these settings.

Ethical Considerations

MultiClinSum consists of published, de-identified clinical case reports and is distributed under the Creative Commons Attribution 4.0 International licence. Llama 3.1 is used under the Llama 3.1 Community License Agreement, LLaDA under the MIT License, and UMLS resources under the UMLS Metathesaurus License Agreement. The models and outputs are research prototypes and are not intended for clinical use without expert review, and the LLM judges are not a substitute for clinician evaluation.

References

- Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 2–30.
- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): integrating biomedical terminology](#). *Nucleic Acids Research*, 32(suppl_1):D267–D270.
- Richard H Browne. 1995. On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*.
- Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris Ebby, Jillian Gorski, Matthew Kalscheur, Samy Khalil, Marie Pisani, Tyler Rubeor, Peter Stetson, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. [Development and validation of the provider documentation summarization quality instrument for large language models](#). *Journal of the American Medical Informatics Association*, 32(6):1050–1060.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.
- Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. 2026. LLaDA-MedV: Exploring large language diffusion models for biomedical image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2508.01617.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [QAFactEval: Improved QA-based factual consistency evaluation for summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12280*.
- François Grolleau, Emily Alsentzer, Timothy Keyes, Philip Chung, Akshay Swaminathan, Asad Aali, Jason Hom, Tridu Huynh, Thomas Lew, April S. Liang, Weihang Chu, Natasha Z. Steele, Christina F. Lin, Jingkun Yang, Kameron C. Black, Stephen P. Ma, Fateme N. Haredasht, Nigam H. Shah, Kevin Schulman, and Jonathan H. Chen. 2026. MedFactEval and MedAgentBrief: A framework and workflow for generating and evaluating factual clinical summaries. In *Pacific Symposium on Biocomputing 2026*. ArXiv:2509.05878.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Ruurd Jan Anthonius Kuiper, Lars de Groot, Bram van Es, Maarten van Smeden, and Ayoub Bagheri. 2025. [LAD: LoRA-adapted diffusion](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–110.
- Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and

- Zhiyong Lu. 2016. [BioCreative V CDR task corpus: a resource for chemical disease relation extraction](#). *Database: The Journal of Biological Databases and Curation*, 2016:baw068.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Bing Liu, Lijun Liu, Xiaobing Yang, Wei Peng, and Li Liu. 2026. [Redefining medical visual question answering using conditional generative diffusion models](#). *Biomedical Signal Processing and Control*, 111:108222.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and robust models for biomedical natural language processing](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38. ArXiv:2502.09992.
- Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. 2024. [Large language models in medicine: the potentials and pitfalls: a narrative review](#). *Annals of Internal Medicine*, 177(2):210–220.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. [Med-HALT: Medical domain hallucination test for large language models](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Matt Post and David Vilar. 2018. [Fast lexically constrained decoding with dynamic beam allocation for neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 1314–1324.
- Miguel Rodríguez-Ortega, Eduardo Rodríguez-Lopez, Salvador Lima-López, Carlos Escolano, Maite Melero, Lorenzo Pratesi, Laura Vigil-Giménez, Letícia Fernandez, Eulàlia Farré-Maduell, and Martin Krallinger. 2025. Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*. Dataset: doi:10.5281/zenodo.17341582.
- Julius Sim and Martyn Lewis. 2012. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *Journal of Clinical Epidemiology*, 65(3):301–308.
- Christine Sinsky, Lacey Colligan, Ling Li, Martha Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.
- Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.
- Liyan Tang, Zhaoyi Sun, Betina Iday, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. [Evaluating large language models on medical evidence summarization](#). *npj Digital Medicine*, 6(1):158.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoît Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*, 30(4):1134–1142.
- Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.
- Lawrence L. Weed. 1968. Medical records that guide and teach. *New England Journal of Medicine*, 278(11):593–600.
- Kevin Yang and Dan Klein. 2021. [FUDGE: Controlled text generation with future discriminators](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

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

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. [Learning to summarize radiology findings](#). In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*. ArXiv:2306.05685.

Appendix

A Implementation Details

Software Dependencies

LLaDA 8B requires transformers==4.38.2 due to upstream dependency constraints in its custom attention implementation. All other dependencies use current stable releases.

Code Availability

The training, inference, and Entity-Pinned Decoding code is available at <https://github.com/bagheria/DiffuClinic>.

B Training Configurations

Table 3 gives the fine-tuning configuration for each model. The LoRA settings, optimiser, and schedule are the same for both. The models differ in the loss each paradigm uses. Zero-shot inference uses the Instruct checkpoints, since the base checkpoints without instruction tuning did not produce usable zero-shot output. The fine-tuned rows use the base checkpoints with LoRA.

Table 4 lists the Entity-Pinned Decoding settings used at inference.

At inference, Llama 3.1 decodes greedily with a cap of 128 new tokens, one document at a time. LLaDA without EPD uses the generation length, step count, block length and remasking of Table 4, with temperature 0, and enables the dLLM-Cache feature cache.

Setting	Llama / LLaDA
FT base ckpt	Base (8B)
ZS ckpt	Instruct (8B)
LoRA rank r	8
LoRA α	32
Dropout	0.1
Targets	q,k,v,o_proj
Epochs	3
Batch size	1
Grad accum	8
Learning rate	2×10^{-4}
Weight decay	0.01
Warmup steps	20
Grad clip	1.0
Precision	bfloat16
Loss span	summary only*

Table 3: Fine-tuning configuration. Both are 8B-parameter models. *Llama uses causal LM loss over summary tokens, LLaDA uses the masked-diffusion objective over summary tokens.

EPD parameter	Value
Generation length L	128
Denoising steps	64
Mask token ID	126336
$\lambda_{\max}$ (logit bonus)	2.0
Annealing power	1.0
Remasking	low-confidence
Block length	16
Budget fraction	0.30
$K_{\text{med}} / K_{\text{dis}}$	4 / 8
Recency threshold / mult.	0.75 / $2\times$
Max entity tokens	10
IDF smoothing	1.0 (Laplace)

Table 4: Entity-Pinned Decoding configuration.

C Evaluation Metric Configurations

C.1 SummaC (Tier 1)

SummaC-ZS, sentence-level granularity, NLI backbone `mnli` (RoBERTa-large-MNLI), batch size 32. Device: CUDA (A10 24 GB). Score range approx. $[-1, +1]$, negative = contradiction dominates entailment.

C.2 QAFactEval (Tier 1)

Question valhalla/t5-base-qg-h1. Generation: Question Answering: `deepset/roberta-base-squad2`. Max 3 questions per summary, QA confidence threshold 0.1. Device: CUDA.

Questions are generated from the source note. The score is the mean, over the questions, of a per-question value equal to 1 when the answer spans extracted from the source note and from the summary are identical after lowercasing, 0 when they differ, and 0.5 when either passage yields no answer above the confidence threshold. A document for which no question is generated scores 0.5, so scores lie in $[0, 1]$. We refer to this metric as QAFactEval for brevity. It is not the implementation of Fabbri et al. (2022).

C.3 MedNER-F1 (Tier 1)

NER model: `en_ner_bc5cdr_md` (scispaCy). Entity types: Disease and Chemical (BC5CDR labels, no UMLS filtering). Precision, recall and F1 are computed per document between the lowercased spans extracted from the generated summary and from the source note, and averaged over the 175 documents.

C.4 Tier 2: Linguistic Similarity

ROUGE-1/2/L	<code>rouge_score</code> (Google Research)
BLEU-1/2/3/4	NLTK, SmoothingFunction method 1
METEOR	NLTK <code>meteor_score</code>
BERTScore-F1	<code>bert-score</code> , roberta-large, rescaled with baseline

C.5 PDSQI-9 LLM Judge Prompt

You are an expert clinical text evaluator. Your task is to evaluate a clinical summary generated by an AI model against the source clinical note. Use the PDSQI-9 rubric below. Score each factor on the specified 1–5 scale. Base all judgments strictly on the source note, not general medical knowledge. Output a JSON object with one integer score per factor. Do not include explanations in the output.

The prompt is reproduced as sent. Its reference to a 1–5 scale does not match the 0–1 items and the NA rule in the rubric that followed it (Appendix D).

The full rubric and scoring anchors are given in Appendix D.

C.6 Judge Configuration

	o3-mini	DeepSeek-R1
Provider	OpenAI API	DeepSeek API
Temperature	0	0
Max tokens	500	500

Both judges evaluated every summary in the test subset. The two judges are drawn from distinct model families, a commercially closed model and an open-weight model, as a check against single-evaluator bias. They agree on the Overall ranking except for one adjacent pair (Section 4).

D Rubric Adapted from the PDSQI-9

Factor	Score	Anchors
Accuracy	1–5	1 = Fabrication. 3 = Misaligned context. 5 = All traceable.
Completeness	1–5	1 = Multiple pertinent omissions. 3 = One. 5 = None.
Organisation	1–5	1 = Incoherent order. 3 = Same as input. 5 = Fully organised.
Synthesis	1–5	1 = Wrong reasoning. 3 = No synthesis. 5 = Integrated synopsis.
Overall	1–5	Holistic clinical utility.
Usefulness	1–5	1 = Nothing pertinent. 5 = All pertinent.
Comprehensibility	1–5	1 = Overly complex. 5 = Plain language.
Succinctness	1–5	1 = Wordy, redundant. 5 = Fewest words.
Citations	1–5	1 = None/incorrect. 5 = All correct, prioritised.
Abstraction	0–1	0 = Not needed. 1 = Needed.
Stigmatisation (summ.)	0–1	0 = None. 1 = Present.
Stigmatisation (source)	0–1	0 = None. 1 = Present.

When Abstraction = 0, Synthesis is scored as NA.

E Qualitative Examples

Two representative samples are presented below.

Example 1: Vaginal Wall Mass (Sample 1944)

Reference: A 41-year-old female was presented with urinary symptoms. There was history of caesarean section. Physical examination revealed an anterior vaginal wall mass. Pelvic MRI showed an inter vesico-vaginal mass, suggesting a leiomyoma. Surgical excision was performed by the vaginal route. There were no postoperative complications. Histopathology examination showed focal endometriosis.

Type	Source note	Reference summary	summary
Gender	"29-year-old <i>female</i> patient ... diagnosed with influenza B."	"29-year-old <i>male</i> emergency department with abdominal pain."	
Age	"An <i>69</i> -year-old man with a history of kidney failure was admitted to our hospital."	"We report the case of an <i>81</i> -year-old man with heterochronous quadruple EMPD."	
Medication	"...recovered with drainage, <i>antibiotic therapy</i> and invasive mechanical ventilation."	"...consistent with bacterial meningitis, and <i>ceftriaxone</i> was initiated."	

Table 5: Examples of entity inconsistencies in Multi-ClinSum GS reference summaries across three inconsistency types. In each case the reference summary conflicts with the source note on a clinically relevant entity.

Llama ZS. A 41-year-old woman, gravida 4, para 2, with a history of 2 abortions and a previous C-section, presented with symptoms of dysuria, pollakiuria, pelvic heaviness, and acute urinary retention. Physical examination and imaging studies (ultrasound and MRI) revealed a 60×40×38 mm intervesical mass, later confirmed to be a cervical leiomyoma. The mass was surgically resected via a vaginal approach. . .

Llama LoRA. A 41-year-old woman, 4 gravida, 2 para, 2 Abortions with a history of C-section delivery. The patient presented with dysuria, pollakiuria, pelvic heaviness and an episode of acute urinary retention. Physical examination revealed a solid, well limited and fixed mass on the anterior vaginal wall. Pelvic ultrasound revealed a heterogeneous inter vesico-vaginal mass measuring 60/40 mm. . .

LLaDA ZS. A 41-year-old woman with a history of 4 deliveries and 2 abortions presented with dysuria, pollakiuria, pelvic heaviness, and acute urinary retention. Physical examination revealed a mass mass on the anterior vaginal wall, and imaging studies showed a heterogeneous intervesico-vaginal mass. MRI and ultrasound suggested a cervical leiomyoma. . .

LLaDA LoRA. A 41-year-old woman, 4 gravida, 2 para, 2 Abortions with a history of C-section delivery. The patient presented with dysuria, pollakiuria, pelvic heaviness and an episode of acute urinary retention. Physical examination revealed a solid, well limited and fixed mass on the anterior vaginal wall. Pelvic ultrasound showed a heterogeneous inter vesico-vaginal mass measuring 60/40 mm. MRI showed a. . .

LLaDA EPD. A 41-year-old woman, 4 gravida, 2 para, 2 Abortions with a history of C-section delivery. The patient presented with dysuria, pollakiuria, pelvic heaviness and an episode of acute urinary retention. Physical examination revealed a solid, well limited and fixed mass on the anterior vaginal wall. Pelvic ultrasound revealed a heterogeneous inter vesico-vaginal mass measuring 60/40 mm. . .

Observations. Fine-tuned models produce nearly identical, extractive summaries, consistent with the statistical indistinguishability on entity-level factuality (Section 4.2.1).

Example 2: Adult Tethered Cord Syndrome (Sample 2522)

Reference: A 47-year-old Asian woman was admitted to our hospital with a 1-year history of worsening leg numbness and urinary dysfunction. Magnetic resonance imaging revealed a low-lying conus medullaris extending to the level of S2 and surrounded by fat tissue at that level. We diagnosed her condition as adult tethered cord syndrome, and spine-shortening vertebral osteotomy was planned. . .

Llama ZS. *Case Report: Adult Tethered Cord Syndrome Treated with Spine-Shortening Vertebral Osteotomy.* A 47-year-old Asian woman presented with a 1-year history of worsening leg numbness and urinary dysfunction. She had a 5-year history of bilateral toe numbness, which progressed to involve her entire toes. Magnetic resonance imaging revealed a low-lying conus medullaris extending to the level of S2. . .

Llama LoRA. A 47-year-old Asian woman was admitted to our hospital with a 1-year history of worsening leg numbness and urinary dysfunction. She first noticed her bilateral 3rd and 4th toe numbness for 5 years. Her symptoms worsened, and her whole toes became numb a year ago. She had also been experiencing frequent urination for a year. . .

LLaDA ZS. A 47-year-old Asian woman with a history of worsening leg numbness and urinary dysfunction was diagnosed with adult tethered cord syndrome. Magnetic resonance imaging revealed a low-lying conus medullaris extending to the level of level S2. The shortening osteotomy was performed at the L2 level, with bilateral spinal screws placed at L1 and L3. . .

LLaDA LoRA. A 47-year-old Asian woman was admitted to our hospital with a 1-year history of worsening leg numbness and urinary dysfunction. The first examination, magnetic resonance imaging, revealed a low-lying conus medullaris extending to the level of S2 and surrounded by fat tissue. 2. Weening vertebral osteotomy was performed. the patient condition was uneventful. the postoperative course was uneventful.

LLaDA EPD. A 47-year-old Asian woman was admitted to our hospital with a 1-year history of worsening leg numbness and urinary dysfunction. She first noticed her bilateral 3rd and 4th toe numbness for 5 years. Her symptoms worsened, and her whole toes became numb a year ago. She had also been experiencing frequent urination for a year. . .

Observations. Llama ZS adds a structured heading not present in the reference (zero-shot structural elaboration). LLaDA LoRA shows minor disfluency ("Weening," and "was uneventful" repeated across two sentences).

F Sample Size and Power Calculations

To determine the required sample size for our evaluation without resorting to arbitrary subsampling, we conducted a pilot-guided power analysis. The calculation was driven by two statistical constraints: ensuring a precise 95% confidence interval (CI) and maintaining sufficient statistical power to detect meaningful differences between models.

1. Sample Size for Confidence Interval Precision

Based on our pilot evaluation of 50 samples, the largest standard deviation among the Tier 1 metrics was $\sigma = 0.269$ (SummaC). To guarantee a 95% CI with a half-width (margin of error) of $E = 0.04$, we calculate the required sample size n using the standard formula:

$$n = \left(\frac{Z_{1-\alpha/2} \cdot \sigma}{E} \right)^2$$

Where $Z_{1-\alpha/2} = 1.96$ (the critical value for a 95% confidence level), $\sigma = 0.269$, and $E = 0.04$. Substituting these values:

$$n = \left(\frac{1.96 \cdot 0.269}{0.04} \right)^2 \approx (13.181)^2 \approx 173.74$$

Rounding up to the nearest whole number yields a minimum requirement of 174 samples. We selected a slightly conservative sample size of $n = 175$.

2. Verification of Statistical Power We subsequently verified the statistical power ($1 - \beta$) to detect an effect size (minimum detectable difference) of $\Delta = 0.075$ in a paired evaluation setup. The required sample size for a target power of 80% is calculated as:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot \sigma_d^2}{\Delta^2}$$

Using the standard normal deviates $Z_{0.975} = 1.96$ ($\alpha = 0.05$) and $Z_{0.80} = 0.8416$ ($\beta = 0.20$), and conservatively assuming the standard deviation of differences σ_d approximates our maximum observed $\sigma = 0.269$:

$$\begin{aligned} n &= \frac{(1.96 + 0.8416)^2 \cdot (0.269)^2}{(0.075)^2} \\ &\approx \frac{7.849 \cdot 0.07236}{0.005625} \approx 100.97 \end{aligned}$$

Because our chosen sample size of 175 comfortably exceeds the 101 samples required for an 80% target, it provides an actual statistical power well above 80% for detecting an effect size of 0.075.

3. Pairwise Tests Pairwise tests use a paired bootstrap of the mean difference with the same resampling. Two-sided p -values are the doubled proportion of bootstrap differences on the far side of zero.