NARRATIVE FIDELITY: BEYOND FACTUALITY IN EVALUATING CLINICAL DIALOGUE SUMMARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical notes are usually evaluated as summaries, with ROUGE, BERTScore, or other overlap with a reference note. These metrics reward factual completeness but say little about whether a note preserves the patient's own framing, concerns, uncertainty, and perspective. We call this property *narrative fidelity* and introduce a six-dimension annotation taxonomy for it. We apply the taxonomy to notes from six systems on the ACI-Bench clinical dialogue corpus, and report two findings. First, narrative fidelity agrees with ROUGE on system ranking (Spearman $\rho = 0.94$) but not on individual cases. Across cases and systems, its correlation with ROUGE-1 is moderate ($r \approx 0.43$–0.68 across core dimensions), well below the correlation among the ROUGE variants themselves ($r \approx 0.76$–0.88), so the taxonomy captures variance that ROUGE does not. ROUGE tracks Clinical/Factual Content most closely ($r \approx 0.68$) and Uncertainty/Epistemic Stance least ($r \approx 0.43$). BERTScore correlates less still ($r \approx 0.19$–0.27 per core dimension) and barely separates the systems: mean BERTScore varies by 0.014 across systems whose narrative fidelity ranges from 0.243 to 0.965. Second, we test sensitivity to a naturally occurring degradation, comparing ASR and human transcripts of the same 28 encounters. Aggregate narrative fidelity and ROUGE-1 change little, but Patient Concerns & Priorities declines ($p = 0.014$, paired t-test; $p = 0.018$, Wilcoxon), and the decline is spread across cases rather than driven by outliers. This result does not survive correction for the six dimensions tested, so we treat it as a hypothesis for replication. It suggests that transcript quality can erode specific narrative content when overall note quality looks unchanged. Inter-annotator agreement is strong on three dimensions and lower for Perspective/Attribution, and we discuss what these results imply for evaluating AI scribes beyond factuality checks.

1 INTRODUCTION

AI systems that turn doctor–patient conversations into clinical notes are usually evaluated the way summarization systems have long been evaluated: compare the generated text to a reference note with ROUGE (Lin, 2004) or BERTScore (Zhang et al., 2020). These metrics ask whether the right facts show up in roughly the right words. They don't ask whether the note preserves *how* the patient told their story – what worried them, how certain they were, whose perspective a given statement belongs to. We call this second property **narrative fidelity**. It's conceptually distinct from factual completeness, and it matters clinically: a note can be factually complete and still flatten a patient's stated uncertainty into false certainty, or swap their stated priority for the clinician's framing. That note is worse than its ROUGE score would suggest.

This paper makes three contributions. The first is a six-dimension, five-status annotation taxonomy for narrative fidelity in clinical note generation (Section 3.3), built to be usable by non-clinician annotators working from a written rubric and worked examples. The second is an empirical demonstration, across six generation systems and dialogues from ACI-Bench (Yim et al., 2023), that narrative fidelity is measurably distinct from standard automatic metrics - correlated with them, but not reducible to them, with the distinction concentrated in specific dimensions rather than spread evenly across the taxonomy (Section 4). The third is a case study built on real, naturally oc-

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.

curring transcript-quality degradation: ASR-quality transcripts produce a detectable, dimension-specific narrative-fidelity loss that an aggregate score or a standard automatic metric misses entirely (Section 4.3).

We frame this as a contribution to evaluation methodology, not a systems paper. We aren't proposing a new note-generation system but just a way to measure a failure mode that current evaluation practice doesn't surface, and a worked example of why that failure mode matters.

2 Related Work

Clinical note generation and AI scribes. A growing line of work applies large language models to generate structured clinical notes from doctor–patient dialogue. The MEDIQA-Chat shared tasks (Ben Abacha et al., 2023a) set a standard benchmark for this problem, alongside the MTS-Dialog corpus of doctor–patient conversations paired with clinical notes (Ben Abacha et al., 2023b). ACI-Bench (Yim et al., 2023), which we build on directly, extends this setting with multiple transcript-quality variants of the same encounter. All of this sits within a broader tradition of dialogue summarization on non-clinical corpora – SAMSum (Gliwa et al., 2019), DialogSum (Chen et al., 2021) – which set the core dialogue-to-summary task formulation and the evaluation conventions clinical note generation has largely inherited, n-gram and embedding overlap against a single reference included.

Automatic summarization and generation evaluation. ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) are still the dominant n-gram overlap metrics for summarization and generation quality. BERTScore (Zhang et al., 2020) scores similarity with contextual embeddings from BERT (Devlin et al., 2019); BARTScore (Yuan et al., 2021) extends this further into likelihood-based similarity. Meta-evaluations – SummEval (Fabbri et al., 2021), the extreme-summarization benchmark XSum (Narayan et al., 2018) – have shown repeatedly that these metrics correlate only moderately with human quality judgments, and specifically fail to separate informativeness from faithfulness. That's the gap our taxonomy targets, though along axes – perspective, concern, uncertainty – that prior faithfulness-focused metrics don't attempt to capture.

Factual consistency and hallucination. A substantial literature studies factual consistency in abstractive summarization specifically. Cao et al. (2018) and Kryściński et al. (2020) propose fact-aware generation and classifier-based consistency checking, respectively; Maynez et al. (2020) run a systematic human annotation study showing that most abstractive summaries on XSum contain some hallucination; Goyal & Durrett (2021) refine this into a fine-grained factuality taxonomy; Fabbri et al. (2022) improve QA-based consistency evaluation. Ji et al. (2023) survey the area broadly. Our taxonomy complements this line of work rather than competing with it: factual consistency asks whether a generated claim is *true*; narrative fidelity asks whether a *true* claim kept its original framing, certainty, and attribution. Our Clinical/Factual Content dimension alone doesn't capture that distinction – Patient Concerns & Priorities and Perspective/Attribution are what's designed to.

LLM-based evaluation. Recent work uses large language models themselves as evaluators – G-Eval (Liu et al., 2023), pairwise LLM-judge protocols validated against human preference data (Zheng et al., 2023). We treat an LLM-judge pass against our human-annotated taxonomy as a natural extension of this work (Section 5), not a replacement for the human annotation reported here; LLM judges have well-known limitations on expert-domain tasks.

Perspective, stance, and epistemic stance in text. Our Perspective/Attribution and Uncertainty/Epistemic Stance dimensions draw on two established NLP subfields: stance detection (surveyed by Küçük & Can (2020)), which studies whether a text takes a position relative to a target, and hedging/uncertainty detection (a shared task since Farkas et al. (2010)), which studies linguistic markers of epistemic commitment. As far as we know, neither has previously been adapted into an evaluation taxonomy for AI-generated clinical documentation.

Annotation reliability methodology. Our reliability diagnosis in Section 3.4 connects to broader work on inter-annotator agreement – Artstein & Poesio (2008)'s survey of agreement coefficients for computational linguistics, and the categorical-agreement statistics of Cohen (1960), Fleiss (1971), and Landis & Koch (1977). Our finding that Pearson correlation can mislead when one annotator's scores have near-zero variance is a concrete case of a general warning in that literature: an agreement statistic built for one data-generating process (balanced categorical ratings, say) can behave badly

under another (a heavily skewed rating distribution). We'd recommend pairing correlation-based IAA with a raw agreement check for narrative annotation tasks like this one.

3 Data and Methods

3.1 Corpus

We use ACI-Bench (Yim et al., 2023), a public corpus of 207 clinical dialogue–note pairs across three sub-collections (*aci*, *virtassist*, *virtscribe*), with professionally authored gold notes held fixed across transcript-quality variants of the same case. One sub-collection, *virtscribe*, separately provides an ASR transcript and a human transcript of the same encounter, both attached to the same gold note – a naturally occurring, non-synthetic degradation axis we exploit in Section 4.3 ($n = 28$ paired cases with both variants present). Raw-text word error rate between the two transcript types is roughly 21–25%, driven mostly by missing speaker tags and punctuation in the raw ASR output; content-word WER, stripping tags and punctuation, is roughly 6–11%. We checked by direct comparison that gold notes are identical across transcript-quality variants for every case – the right design for isolating whether transcript quality affects generated-note quality independent of the reference.

3.2 Systems

We compare six note-generation systems spanning vendors and capability tiers, all built on transformer-based architectures (Vaswani et al., 2017): two current Claude models (Claude Sonnet 5 and Claude Haiku 4.5), generating notes under one fixed prompt we control, and four published baselines reused from the original ACI-Bench challenge under its CC BY 4.0 license, each covering the same test splits – GPT-4 (OpenAI, 2023), a ChatGPT-era model, `text-davinci-003` (an instruction-tuned GPT-3-family model in the lineage of Brown et al., 2020; Ouyang et al., 2022), and a BART-large baseline (Lewis et al., 2020). Reusing the published baselines lets us compare across vendors without needing API access beyond a single provider, and keeps the reference notes and scoring procedure identical across all six systems. For the transcript-quality experiment (Section 4.3) we use Claude Haiku 4.5 only, generating notes from both the human-transcribed and ASR-transcribed version of each encounter.

3.3 Narrative Fidelity Taxonomy

We built a two-axis annotation scheme: six content dimensions crossed with five transformation statuses. The *core tier* - Clinical/Factual Content, Patient Concerns & Priorities, Uncertainty/Epistemic Stance, Perspective/Attribution - carries our headline claims; an *extended tier*, Emotional Tone and Context/Social History, is reported alongside it but doesn't. For each dimension, an annotator identifies the narrative units present in the source dialogue and marks each as Preserved, Weakened, Distorted, Omitted, or Introduced (hallucinated) in the generated note relative to that dialogue. A dimension score is computed as

$$\text{score} = \frac{\#\text{Preserved} + 0.5 \times \#\text{Weakened}}{\#\text{scored units}}, \tag{1}$$

Introduced units are tracked separately as a hallucination rate rather than folded into this score. The full written rubric – decision rules, worked examples included is in the supplementary material.

3.4 Annotation Process and Reliability

Three annotators, none involved in designing the taxonomy, applied the rubric to 117 cases stratified across the corpus's three sub-collections, each scored against all six systems (702 case–system judgments). A 22-case subset was independently annotated by all three annotators to assess reliability; the remaining 95 cases were divided roughly evenly among them. We report agreement on the triple-annotated subset using Krippendorff's alpha (interval level), which accommodates more than two raters and incomplete rating coverage. Pooled across the 22 cases, agreement is strong for Clinical/Factual Content ($\alpha = 0.951$), Context/Social History ($\alpha = 0.958$), and Patient Concerns & Priorities ($\alpha = 0.836$), and moderate for Uncertainty/Epistemic Stance ($\alpha = 0.771$) and Emotional

Table 1: System-level core-tier narrative fidelity (NF-core), mean across annotators, $n = 117$ cases per system.

System	NF-core (mean)
Claude Sonnet 5	0.965
Claude Haiku 4.5	0.893
GPT-4 (baseline)	0.582
ChatGPT (baseline)	0.525
`text-davinci-003` (baseline)	0.486
BART-large (baseline)	0.243

Tone ($\alpha = 0.686$). Perspective/Attribution shows lower agreement ($\alpha = 0.417$), and we read its absolute scores as indicative (Section 5). We report pooled rather than per-system alpha: conditioning on which units each system's output contains leaves as few as 2–4 triple-annotated cases per system, and under low rating variance chance correction artificially depresses alpha for at least two systems despite high raw agreement (Section 4.4).

In the two-annotator ASR experiment (Section 4.3), Clinical/Factual Content shows a near-zero Pearson correlation between annotators ($r = -0.01$) that reflects a calibration difference rather than disagreement. On the 448 units with exactly matching text, 80% were rated PRESERVED by both annotators. One annotator scored nearly all units as fully preserved (mean 0.998, s.d. 0.014), while the other showed more spread (mean 0.842, s.d. 0.140), which leaves little variance for a correlation to detect. We therefore report raw agreement alongside correlation-based measures.

As an external check on annotation validity, an independent clinician reviewer (a resident physician uninvolved in taxonomy design or annotation) ran a recognition-task spot-check. Rather than re-segmenting each dialogue from scratch, which reintroduces the segmentation mismatch that limits agreement between independent annotators, the reviewer judged whether they agreed with an already-assigned transformation status. The sample was 62 narrative-unit judgments across all four core dimensions and both experiments: 34 units from two main-comparison cases scored against three systems spanning the fidelity range, and 28 units from two ASR pairs. The reviewer agreed with 53 of 62 (85.5%; Clinical/Factual Content 87.5%, $n = 48$; Patient Concerns & Priorities 85.7%, $n = 7$; Perspective/Attribution 75.0%, $n = 4$; Uncertainty/Epistemic Stance 66.7%, $n = 3$). The one disagreement on the ASR pairs concerned a Patient Concerns & Priorities judgment on the ASR side of *test1:VS000*, which the reviewer read as PRESERVED rather than WEAKENED; under that correction the pair shows no change instead of a drop. The other three Patient Concerns judgments sampled, including both sides of *test3:VS030*, were confirmed. The check covers 2 of the 27 scorable pairs, so it complements rather than replaces full re-annotation.

4 RESULTS

4.1 SYSTEM-LEVEL COMPARISON

Table 1 reports mean core-tier narrative fidelity by system, pooled across the full annotated case set: 702 case–system judgments, $n = 117$ cases per system, averaged across annotators wherever more than one scored a case.

4.2 DISTINCTNESS FROM AUTOMATIC METRICS

Narrative fidelity agrees with established metrics where they should agree, and adds information where they cannot. At the system level, narrative fidelity and ROUGE-1 rank the six systems almost identically (Spearman $\rho = 0.94$), which supports the taxonomy as a valid measure of generation quality. At the case level, a mixed-effects model (Bates et al., 2015) with case as a random intercept, treating the 6 systems $\times$ 117 cases as 702 non-independent observations, shows that within a case, higher ROUGE robustly predicts higher narrative fidelity ($p < 0.0001$).

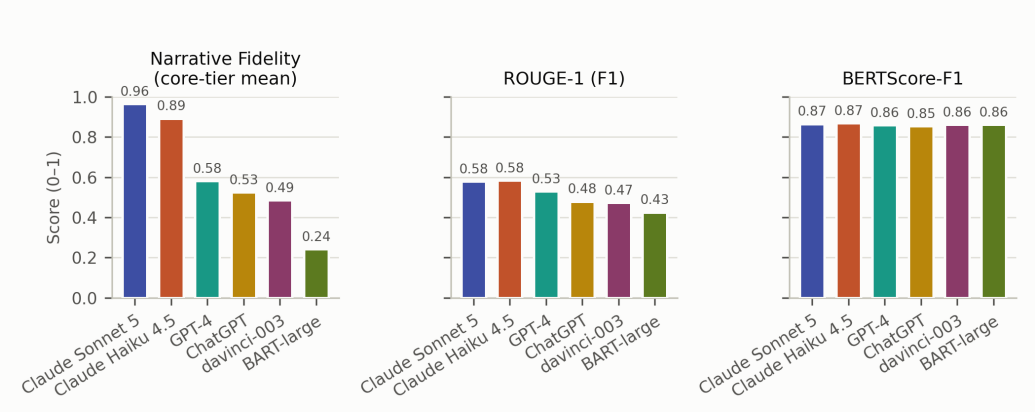

Figure 1: System-level mean narrative fidelity, ROUGE-1, and BERTScore-F1 across the six systems ($n = 117$), on a shared 0–1 axis. Narrative fidelity spans the full range; ROUGE-1 varies moderately; BERTScore-F1 is nearly flat.

Table 2: Correlation between automatic metrics and narrative fidelity, pooled across systems ($n = 696$ case–system pairs unless noted), alongside each metric's own internal or cross-metric correlation.

Comparison	**Pearson** r	n
ROUGE-1 vs. ROUGE-2/L (internal)	0.76–0.88	696
BERTScore vs. ROUGE-1/2/L (cross-metric)	0.63–0.72	90
ROUGE-1 vs. NF core-tier mean	0.64	696
ROUGE-1 vs. Clinical/Factual Content	0.68	696
ROUGE-1 vs. Uncertainty/Epistemic Stance	0.43	72
BERTScore vs. NF core-tier mean	0.29	90
BERTScore vs. Clinical/Factual Content	0.19	90
BERTScore vs. Uncertainty/Epistemic Stance	0.24	42

The relationship is strong enough to validate the measure and distinct enough to show it adds something. Narrative fidelity correlates with ROUGE-1 at $r \approx 0.43$–0.68 across core dimensions (Table 2), well below the $r \approx 0.76$–0.88 at which the ROUGE-1/2/L family correlates with itself, so the taxonomy captures variance that lexical-overlap metrics do not share. BERTScore (Zhang et al., 2020) tracks it less closely ($r \approx 0.19$–0.27 across the four individual core dimensions; $r = 0.29$ against the pooled core-tier mean), even though it is designed to be more sensitive to meaning-preserving paraphrase than n-gram overlap. The difference is clearest at the system level. Mean BERTScore-F1 spans only 0.855 to 0.869 across the six systems, a range of 0.014, and ROUGE-1 spans 0.426 to 0.585. Narrative fidelity spans 0.243 to 0.965, so it separates systems that both automatic metrics group closely together. Among core dimensions, Clinical/Factual Content tracks ROUGE-1 most closely ($r \approx 0.68$), as expected for the dimension nearest to lexical overlap, while Uncertainty/Epistemic Stance tracks it least ($r \approx 0.43$), consistent with that dimension capturing information beyond surface wording.

4.3 Real Transcript-Quality Degradation

We compare 28 paired *virtscribe* cases, each with a human transcript and an ASR transcript of the same encounter, scored against the same gold note, with notes generated by Claude Haiku 4.5 from both. At the aggregate level, core-tier narrative fidelity changes little under ASR transcription, and ROUGE-1 is essentially flat (Table 3). An aggregate-only evaluation would therefore conclude that transcript quality does not matter for note quality.

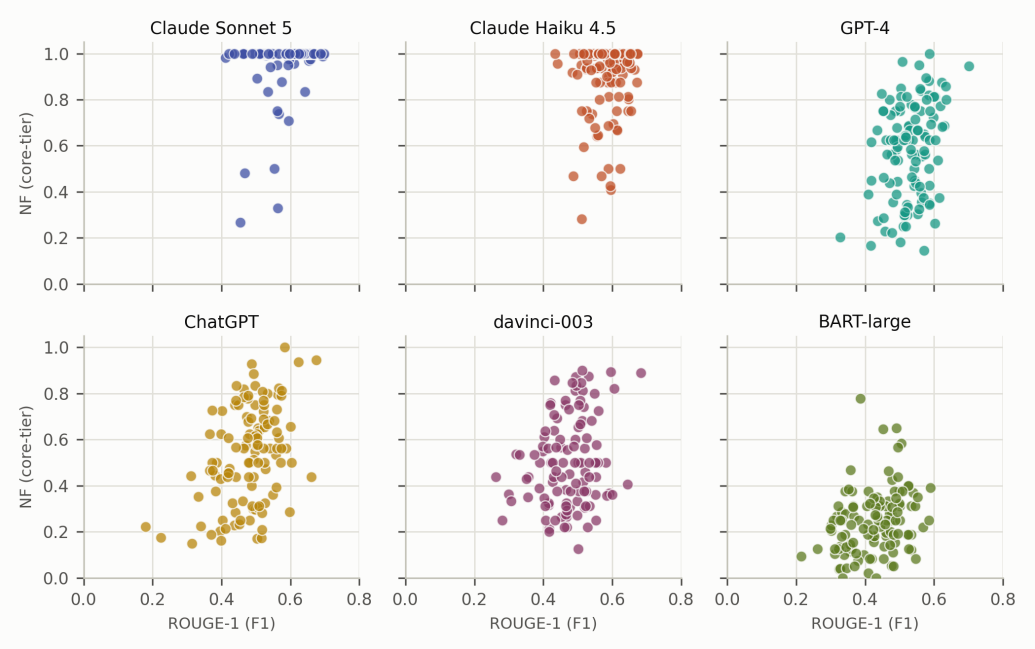

Figure 2: Case-level ROUGE-1 versus narrative fidelity (core-tier mean), faceted by system ($n = 117$ cases per panel). The moderate, noisy relationship (pooled $r \approx 0.64$ for the core-tier mean, Table 2) is visible in every panel, alongside substantial system-level offsets.

Table 3: Narrative fidelity under human transcription versus ASR transcription, 28 paired cases. Significance from paired t-test / Wilcoxon signed-rank test (Wilcoxon, 1945)

Measure	**Human-transcript**	**ASR**	Δ	p (t / **Wilcoxon**)
NF core-tier mean	0.931	0.898	-0.033	0.140 / 0.219
Patient Concerns & Priorities	–	–	-0.107	**0.014 / 0.018**

The dimension-level view shows otherwise. Patient Concerns & Priorities drops under ASR transcription (Table 3), and the drop is not driven by a single case: of the 27 pairs with a scorable score in both conditions, eight decline, one improves, and eighteen are unchanged (Figure 3). The direction matches the mechanism we would expect, since ASR transcription tends to flatten the speaker attribution, hedging, and emphasis that carry patient-stated concerns. The clinical stakes are direct: a patient's stated concern can go unrepresented in their chart note, even while aggregate fidelity and standard automatic metrics show no change. This is the kind of degradation that dimension-aware evaluation surfaces and aggregate metrics miss.

This is the only one of the six dimensions tested to reach nominal significance, and it does not survive Bonferroni correction ($\alpha \approx 0.0083$), so we present it as a hypothesis for replication on a larger paired sample. A clinician spot-check (Section 3.4) also revised one of the sampled Patient Concerns judgments toward no change.

4.4 Inter-Annotator Agreement Summary

Table 4 summarizes inter-annotator agreement across both experiments: pooled Krippendorff's alpha for the three-annotator main comparison (Section 3.4), pairwise Pearson r for the separate two-annotator ASR-degradation experiment. Perspective/Attribution is the least reliable dimension in the main comparison ($\alpha = 0.417$), Emotional Tone next ($\alpha = 0.686$); the remaining four all clear $\alpha = 0.77$. The Clinical/Factual Content near-zero Pearson r in the ASR column is discussed in Section 3.4 and Section 5 – it reflects annotator calibration, not incoherent disagreement.

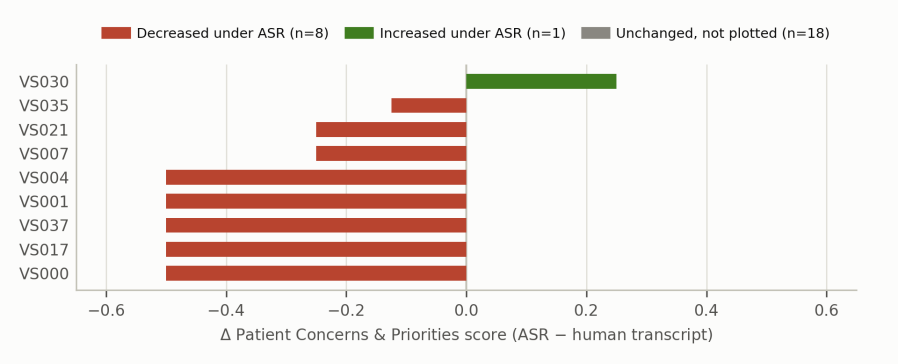

Figure 3: Per-case change in Patient Concerns & Priorities score under ASR transcription versus human transcription, for the 9 of 27 scorable pairs that changed (18 unchanged pairs not plotted; see Section 4.3 for the excluded 28th pair). The drop is distributed across cases of varying magnitude rather than driven by a single outlier.

Table 4: Inter-annotator agreement by dimension. Main comparison: pooled Krippendorff's alpha (interval level of measurement), three annotators, 22-case reliability subset – chosen because it natively handles more than two raters and partial rating coverage, unlike pairwise correlation. ASR experiment: pairwise Pearson r, a separate two-annotator design over 28 cases where a raw-agreement check (Section 3.4) is more informative than a missing-data-tolerant statistic.

Dimension	**Main comparison** (α)	**ASR experiment** (r)
Clinical/Factual Content	0.951	−0.01
Patient Concerns & Priorities	0.836	0.21
Uncertainty/Epistemic Stance	0.771	0.28
Perspective/Attribution	0.417	0.17
Emotional Tone (extended)	0.686	0.05
Context/Social History (extended)	0.958	n/a

5 Discussion and Limitations

The core practical claim here is that evaluating AI-generated clinical notes on factuality or completeness metrics alone, even sophisticated ones - systematically misses certain failure modes, and those failure modes aren't spread evenly across content types. Perspective/attribution loss and patient-concerns erosion under transcription-quality degradation are two concrete cases this taxonomy surfaces that a factuality-only pipeline would have missed.

The six-system comparison (Table 1) reflects existing generation pipelines rather than a controlled ablation: the two models share one prompt designed for this study, while the four baselines are original ACI-Bench challenge outputs produced under their own settings. The paired transcript-quality analysis covers 28 pairs from one sub-collection and a single generation system, so the effect size would benefit from replication. Within it (Section 4.3), Patient Concerns & Priorities is the only dimension to reach nominal significance and does not survive Bonferroni correction, so we treat it as suggestive; 27 of 28 pairs contribute to that dimension.

Reliability for the main comparison (Section 3.4) is established on a 22-case subset annotated by all three annotators. Clinical/Factual Content, Context/Social History, and Patient Concerns & Priorities show strong pooled agreement (Krippendorff's $\alpha \geq 0.84$; Table 4), and Uncertainty/Epistemic Stance is moderate ($\alpha = 0.771$). Perspective/Attribution is lower ($\alpha = 0.417$), so its scores (Figure 5) are best read as indicative, and a refined rubric is a natural next step. The corpus is limited to outpatient primary-care-style encounters in ACI-Bench, so transfer to other clinical settings remains for future work. The taxonomy was designed by one of the authors; none of the three main-comparison annotators took part in that design.

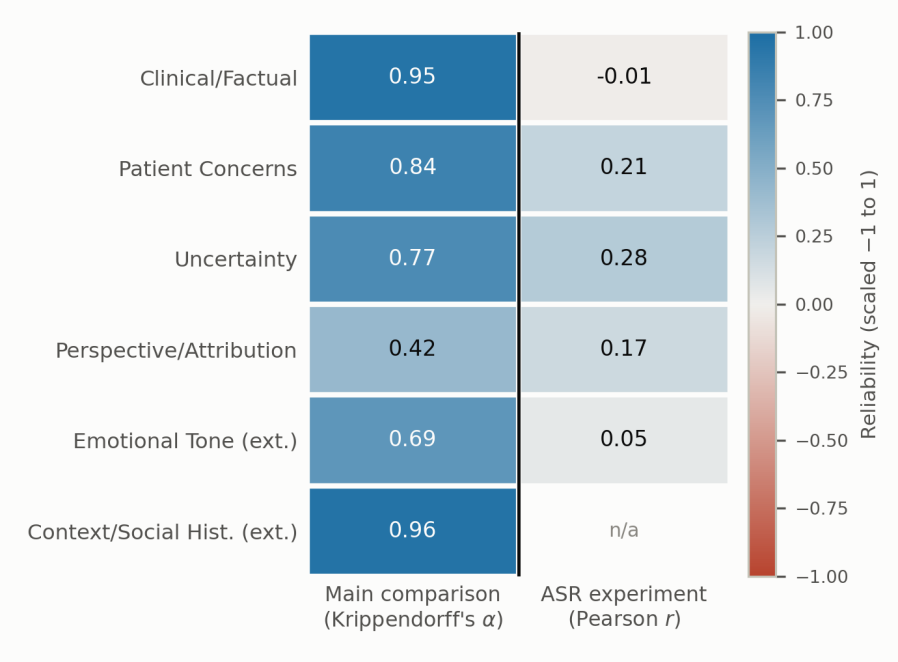

Figure 4: Inter-annotator agreement by dimension: pooled Krippendorff's alpha for the main comparison versus pooled Pearson r for the ASR-degradation experiment (Table 4). Diverging colormap centered at 0. The near-zero Clinical/Factual value in the ASR column reflects an annotator calibration difference, not incoherent disagreement (Section 3.4).

6 Conclusion

We've introduced narrative fidelity as a measurable, distinct dimension of clinical note quality; shown that it diverges from standard automatic metrics in specific, interpretable ways, most clearly on Uncertainty/Epistemic Stance, the core dimension ROUGE tracks worst; and used real transcript-quality variation to show that this distinction has practical consequences an aggregate-only evaluation would miss entirely. We release the rubric and taxonomy as a starting point for evaluation pipelines that look past factual overlap and ask whether a generated note kept faith with the patient's own account of their care.

AI use statement

We used ChatGPT during preparation of this manuscript to draft initial text for portions of the Related Work and Discussion and to edit and polish author-written text for grammar, clarity, and flow. All AI-generated or AI-edited text was reviewed, revised, and verified by the authors, who take full responsibility for the content, accuracy, and citations in this paper.

Ethics statement

This work uses ACI-Bench, a publicly released, de-identified clinical dialogue corpus (Yim et al., 2023, released under its own stated terms for research use). We collected no new patient data; every dialogue and note used here was already de-identified and public before this work started. Our annotation process had multiple annotators labeling AI-generated notes against source dialogues from this public corpus, with no additional human-subjects data collected. Given the clinical domain: our findings are a research question about the reliability of AI-generated documentation, not a validated recommendation for clinical deployment. We don't claim any evaluated system is suitable for unsupervised clinical use.

Reproducibility statement

The narrative fidelity rubric – dimension definitions, decision rules, worked examples – is in the supplementary material, along with the scoring formula from Section 3.3. The annotation pipeline (data preparation, the annotation interface, and the scripts for dimension scores, inter-annotator agreement, and the paired transcript-quality analysis) is included as anonymized supplementary code. The clinician spot-check from Section 3.4 is included in full too – the 62-item review packet, the case context given to the reviewer, the reviewer's completed judgments, and the scoring script. ACI-Bench is publicly available (Section 3.1); the four baseline systems reused in Section 3.2 are the original challenge's published outputs, also public under the corpus's license. The statistical tests we use – paired t-test, Wilcoxon signed-rank test, Pearson/Spearman correlation, Krippendorff's alpha, and the mixed-effects model from Section 4.2 – are standard and run with off-the-shelf statistical libraries, referenced in the supplementary code.

References

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 503–513, Toronto, Canada, 2023a. Association for Computational Linguistics.

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 (EACL)*, 2023b.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 2021.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 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 (NAACL-HLT)*, 2022.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, Uppsala, Sweden, 2010.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, 2019. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248, 2023.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, 2020.

Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys*, 53(1), 2020.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

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

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

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. ACI-BENCH: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10, 2023.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

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. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Appendix

Beyond the four figures in the main text, we include two more here: a full per-dimension breakdown by system, and hallucination (Introduced) rates by dimension. The full rubric and per-case data are in the supplementary material. We don't report a per-system inter-annotator agreement breakdown – at the available sample size (2–4 triple-annotated cases per system on the 22-case reliability subset), per-system Krippendorff's alpha is highly volatile and, for at least two systems, artificially low from chance-correction under low rating variance rather than genuine disagreement (Section 3.4). We report only the pooled, dimension-level estimates in Table 4.

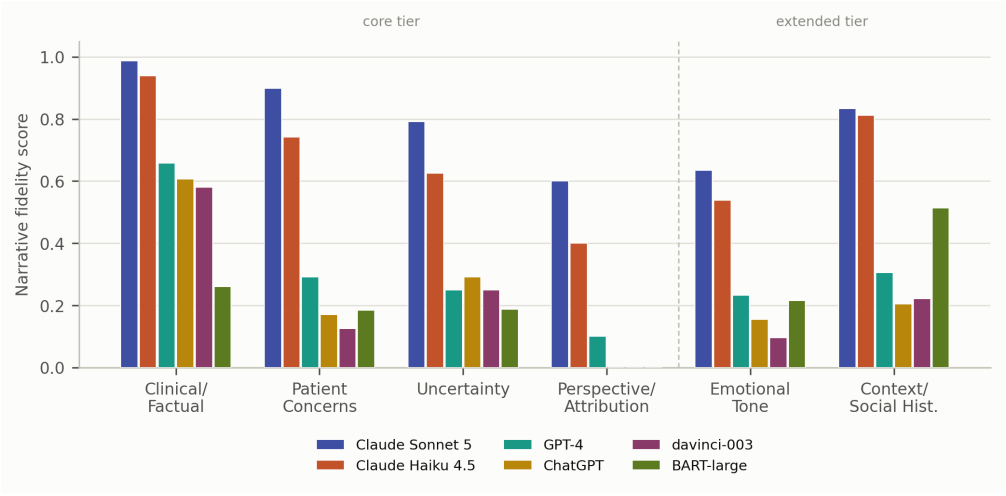

Figure 5: Mean narrative fidelity by system, broken out across all six taxonomy dimensions ($n = 117$ cases per system). Table 1 reports only the core-tier mean; this figure shows the full per-dimension picture, including the two extended-tier dimensions.

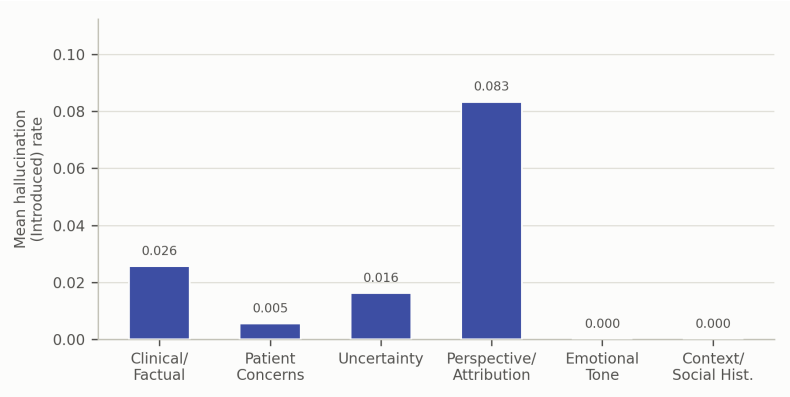

Figure 6: Mean hallucination (INTRODUCED) rate by dimension, pooled across all six systems ($n = 117$ cases per system). Perspective/Attribution shows the highest introduced-content rate (0.083), well above Clinical/Factual Content (0.026); the remaining dimensions are markedly lower still.