Persona-Conditioned Generation of Patient Self-Reports from EHRs

Yuexin Wu[1], **Jianming Wei**[2], **Shiqi Wang**[3], **Vasile Rus**[1]
[1]University of Memphis, Memphis, TN, USA
[2]Central Diagnostics Laboratory, University Medical Center Utrecht, Utrecht, The Netherlands
[3]The Second Clinical College of Guangzhou University of Chinese Medicine, Guangzhou, China
{ywu10, vrus}@memphis.edu, J.wei@umcutrecht.nl, 20251110858@stu.gzucm.edu.cn

Abstract

Accurate diagnosis depends not only on clinical expertise but also on how patients describe their symptoms at first contact. Yet large English corpora of patient-authored self-reports are scarce, limiting advances in natural, context-aware narrative modeling. We address this gap by generating first-person self-reports from structured EHR content conditioned on persona attributes that capture social and clinical context. Reports are produced by two generators and scored by two independent graders using a rubric with four dimensions, complemented by a rubric-free preference test. Across 10k stratified cases, we compare two generators under a reliable evaluation protocol and select the higher-scoring one based primarily on Clinical Correctness and Faithfulness, yielding a dataset composed of narratives from the stronger system. Our contributions are threefold: (I) we developed and release a large, persona-conditioned dataset of patient-style self-reports grounded in patient-stated EHR facts, (II) we introduce a transparent evaluation framework that combines rubric-based scoring with rubric-free preference to mitigate grader bias and enable cross-validation, (III) we find that graders exhibit systematic stylistic preferences in rubric-free approach that influence scores independent of clinical content, and (IV) we study large language models for producing first-person self-reports from structured EHRs, highlighting where they succeed, where they fail, and how this affects use in telemedicine and triage.

Keywords: patient self-report generation, persona-conditioned clinical text, evaluation framework, grader bias and stylistic preference

1. Introduction

Disease diagnosis has become a central pillar of modern healthcare, enabling early detection and timely intervention for acute conditions while guiding lifestyle adjustments and medication regimens to prevent or slow chronic disease.

At first contact, self-reports, i.e., the patients' own narratives, provide the earliest clinically useful evidence. Patient-authored self-reports carry early signals about onset, trajectory, perceived triggers, and functional impact. Self-reports can be collected online before a visit and are well suited to telemedicine, longitudinal follow-up, and population screening (Semigran et al., 2015). They are especially effective for early triage and first-contact routing, and they generalize across front-end applications such as online triage tools, navigation assistants, and conversational agents. Training and evaluating automatic disease diagnosis models on self-reports, as opposed to EHR data, therefore better matches deployment conditions and captures information that categorical checklists or templated notes tend to suppress. Yet large English corpora of such self-reported patient narratives are scarce, which limits progress on natural, context-aware modeling of patient language and the development of more accurate and real-settings automatic diagnosis models (Zeng et al., 2020b).

We address this gap by generating first-person self-reports from structured EHR content while conditioning on persona attributes that reflect social and clinical context. The generation focuses strictly on patient-stated facts and adapts voice and cadence to background cues without introducing new information (Short et al., 2009a; Wynia and Osborn, 2010a). In particular, prior evidence shows that age and gender systematically shape symptom description and presentation (Lautenbacher et al., 2017b; Canto et al., 2000, 2012; van Oosterhout et al., 2020b). To build a reliable resource, we compare two candidate generators under a balanced protocol with two independent graders and select the higher-scoring system with decisions driven primarily by Clinical Correctness and Faithfulness. The resulting dataset contains persona-conditioned narratives produced by the stronger system and grounded in patient-stated evidence.

As already hinted earlier, the proposed first-person self-reports dataset will serve patient-facing applications where concise, faithful narratives matter. Examples include telemedicine intake, online triage, and navigation tools that route patients to appropriate care (Semigran et al., 2015). The resource also enables analysis of how background factors shape first-person reporting (Short et al., 2009a; Wynia and Osborn, 2010a; Berkman et al., 2011b; van Oosterhout et al., 2020b; Lautenbacher et al., 2017b), and how grading practices can introduce stylistic preferences independent of clinical content. In conclusion, our contributions are four fold:

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10792–10801
11–16 May 2026.

- Data: We construct and release a large-scale corpus of persona-aware patient self-reports, tightly aligned with patient-provided clinical facts from electronic records.
- Evaluation: We design a transparent assessment protocol that couples rubric-guided scoring with pairwise preference judgments, enabling bias-aware analysis and cross-rater consistency checks.
- Bias analysis: We discover that graders exhibit systematic stylistic preferences in rubric-free evaluation, revealing how non-clinical stylistic factors can influence perceived report quality independent of clinical correctness.
- Generation: We examine the role of large language models in transforming structured EHR content into first-person self-reports, characterizing their strengths and limitations under persona conditioning and reliable evaluation.

2. Related Work

At the corpus level, related datasets fall into two complementary categories with distinct applications. The first category comprises clinician authored EHR corpora that include admission notes, hospital course, and summaries of laboratory or imaging findings, such as MSDiagnosis (Hou et al., 2024), CMEMR (Jia et al., 2025), and large scale resources like MIMIC (Johnson et al., 2016, 2023); these datasets support document level diagnosis, long form diagnostic inference, and information extraction from comprehensive clinical documentation. The second category comprises patient side resources that collect real online consultations with free text self reports and physician replies, typified by Haodf (Su et al., 2024) and complemented by MedDialog (Zeng et al., 2020a); these datasets are valuable for triage and diagnosis from the patient perspective and for modeling patient facing language, yet they have notable limitations, since Haodf is Chinese only which complicates cross dataset comparison and large scale finetuning, and MedDialog does not provide document level diagnostic gold labels which limits its use for supervised training of diagnosis models.

Complementing corpus design, a broader literature shows that patient narratives are systematically shaped by demographic, socioeconomic, and cognitive factors, which influence both what is reported and how it is phrased. Age affects pain perception, temporal recall, and descriptive detail, with older adults typically reporting fewer or less intense symptoms and providing sparser timelines (Lautenbacher et al., 2017a). Gender also modulates presentation, as women are more likely to express diffuse or atypical sensations during acute cardiac events, whereas men more often produce concise, prototypical accounts (Canto et al., 2007; van Oosterhout et al., 2020a). Self report accuracy varies with population characteristics as well, and large scale validation indicates that younger individuals, males, those with higher education, and healthier respondents report health care utilization and absenteeism more accurately (Short et al., 2009b). Socioeconomic status is tightly linked to communication quality and health literacy, which shape vocabulary, sentence complexity, and the clarity of causal or temporal framing (Wynia and Osborn, 2010b; Verlinde et al., 2012; Berkman et al., 2011a). Taken together, these findings imply that first person clinical narratives are not neutral linguistic artifacts but reflect stable, measurable patterns across age, gender, education, and related socioeconomic factors.

Against this backdrop, our work focuses on generating realistic first person self reports that mirror how patients with diverse backgrounds describe symptoms at first contact. We operationalize empirically supported variables such as age, gender, education, and other contextual cues as persona attributes to shape register, cadence, and selective denials while preserving fidelity to patient stated facts in the EHR. In this way, persona conditioned generation bridges structured clinical evidence and authentic patient voice, complementing prior efforts that emphasize reading and classification rather than patient style narrative writing.

3. Method

Our method proceeds in three stages: (i) data preparation that extracts patient-stated EHR content and persona attributes; (ii) persona-conditioned self-report generation from these inputs; and (iii) rubric-based evaluation with two independent graders.

3.1. Data Preparation

We construct input pairs by extracting patient-stated content from the MIMIC-IV Note corpus and aligning each case with a persona drawn from structured MIMIC-IV tables. The persona includes eight attributes that are available in MIMIC-IV and are empirically linked to how people describe their symptoms: marital status, type of health insurance, primary language, where the encounter began in the system such as emergency department or clinic, whether the visit was emergency, urgent, or scheduled, a hospital coding of case complexity and acuity, gender, and age. Together these cues shape voice without changing facts. Point of entry and visit urgency tend to set cadence and compression

```
SYSTEM_MSG = (
"You are a careful clinical writing assistant. Write exactly one paragraph in the first person ('I…') "
"using only what the patient stated at presentation. Do NOT include exam findings, vitals interpretation, "
"labs/imaging, inpatient course, diagnoses, treatment plans, or past history unless explicitly stated today. "
"Keep explicit denials that the patient stated, but present them briefly and naturally. "
"No headings, no lists, no added facts."
)

PROMPT_TEMPLATE = """Rewrite the patient's self-report as a natural, conversational first-person paragraph.
Adapt tone and word choice to the persona below, while strictly staying within patient-stated information.
Return ONE paragraph in English.

- Persona (JSON):
{persona_json}

- EHR excerpt (patient-stated only):
\"\"\"
{text}
\"\"\"

Denial style rules (to avoid a dense checklist):
- Include ONLY denials explicitly stated in the EHR and relevant to the main concerns.
- Group denials by system and compress into at most TWO short clauses.
- Prefer plain connectors ("I'm not having…", "No…", "I haven't noticed…"); avoid repetitive "I deny…".
- Place denials once after positive symptoms; keep the denial segment brief (~15–25 words).

HPI/PMH inclusion rules (patient-stated only):
- Focus on what's happening today/recently as the patient told it (onset, course, why they came).
- Mention past conditions/procedures/meds ONLY if the patient stated them today and they clarify the complaint.
- No clinician voice (no mechanisms/interpretation/diagnoses/plans or ED course).
- Use patient-time anchors ("today", "yesterday", "over the past two days"); avoid calendar dates unless patient gave one.

Style constraints (derived from persona; do not add facts):
- Education: match vocabulary/sentence length (high/college/master's → concise but fluent; middle → simpler words).
- drg_severity: higher severity → tighter, more focused sentences; lower severity → slightly fuller phrasing.
- Insurance: shape tone, not content; don't mention insurance unless the patient said it.
- Private → calm, organized, confident cadence.
- Medicare → plain, steady, matter-of-fact with clear time anchors.
- Medicaid/Uninsured → slightly urgent, simpler sentences; avoid jargon.
- admission_location / admission_type (only if stated in EHR):
- Emergency → tighter, more focused cadence; minimal elaboration.
- Non-urgent/scheduled/clinic referral → slightly fuller phrasing using only patient-stated context.
- gender & anchor_age: match voice and maturity; avoid slang; age-appropriate simplicity if very old.
- Language: keep English; subtle phrasing influenced by background is okay (no new facts).
- marital_status: minimal day-to-day flavor consistent with background (no new medical facts)."""
```

Figure 1: Prompt used for first-person, patient-stated-only generation.

of the story, with emergency presentations under higher acuity often eliciting brief, time-anchored statements that foreground onset and escalation, whereas scheduled referrals permit fuller context and elaboration. Insurance serves as a practical proxy for socioeconomic context, which relates to health literacy and communication style and thus affects vocabulary and the ordering of details. Primary language influences lexical choice, syntactic simplicity, and preferred expressions, so it guides how plainly or idiomatically a patient narrates the same content. Gender differences are known to influence help seeking and the language used to describe pain or atypical sensations, which alters emphasis and phrasing. Age is tied to sentence length, recall of timelines, and preference for simpler or more varied wording, so it naturally modulates complexity and rhythm. We deliberately do not include attributes that either are not reliably available in MIMIC-IV or would inject clinical conclusions rather than background context, such as race and ethnicity with their ethical and coding concerns, long-term outcomes like discharge disposition or death, and clinician-side evidence including diagnoses, laboratory results, imaging findings, or medication orders; these exclusions avoid information leakage and keep the generation strictly grounded in the patient's voice.

We include education level because it strongly shapes narrative style and vocabulary. As MIMIC-IV does not provide this attribute, we assign each case an age-appropriate category (primary, mid-

```
You are a clinical writing evaluator. Use ONLY the persona and the PATIENT-STATED EHR excerpt as ground truth.
Evaluate the candidate self-report strictly against that.

Inputs
- Persona (JSON): {persona_json}
- EHR (patient-stated only):
{ehr_text}
- Candidate self-report:
{candidate_report}

Scores (1–5) & weights
- ClinicalCorrectness (0.40): Score only the claims actually made in the report (affirmed symptoms or explicit denials).
• Zero-claim rule: if the report makes NO symptom claims at all, set ClinicalCorrectness = 5.0.
• Semantic match: count synonyms/abbreviations/paraphrases if they clearly refer to the same symptom TODAY (e.g.,
SOB ↔ shortness of breath; loose stools ↔ diarrhea; can't keep food down ↔ vomiting).
• Negation scope & timing: a denial is supported only if TODAY'S EHR states the same denial.
• Contradictions: if the report asserts the opposite of TODAY'S EHR, penalize heavily.
• Unsupported new facts: penalize new medical facts not in the EHR; do NOT penalize neutral paraphrases.
• Omissions: do NOT penalize omissions unless they create a contradiction.
- Faithfulness (0.30): Only patient-stated facts; no leakage of exam, vitals interpretation, labs/imaging, inpatient course,
diagnoses, or treatment speculation.
- Persona (0.20): Tone/word choice align with persona cues (education, age, language, setting) without adding facts.
- Realism (0.10): Natural first-person voice; concise; selective denials (not exhaustive lists); plausible length/cadence for
the setting/persona.

overall = 0.30*Faithfulness + 0.40*ClinicalCorrectness + 0.20*Persona + 0.10*Realism

Output (JSON only)
{
"Faithfulness": <1-5>,
"ClinicalCorrectness": <1-5>,
"Persona": <1-5>,
"Realism": <1-5>,
"overall": <1-5>
}
```

Figure 2: Prompt used for evaluation.

dle, high school, college, master's, or PhD) guided by broad U.S. cohort patterns. In practice, age anchors the plausible range (for minors we cap the highest attainable schooling), and we introduce mild randomness around the cohort tendencies so that groups are diverse rather than deterministic. This yields realistic, age-consistent education assignments that influence tone and sentence complexity while leaving clinical content unchanged.

3.2. Generation Pipeline

We generate first-person patient self-reports through a structured prompt-based pipeline that integrates EHR excerpts with persona attributes. To guide the language models during generation, the prompt specifies three complementary rule sets: 1) Denial style rules guide the model in selecting which negative symptoms to mention and how to express them naturally. Because patients in real encounters rarely list many "no" findings, the report includes at most one to two salient denials to keep the narrative focused and realistic, and we expand the set of denials only when the EHR explicitly records them as patient-stated. 2) HPI/PMH inclusion rules guide the model in selecting content related to the history of present illness (HPI) and past medical history (PMH), ensuring that the generated report includes only patient-stated information relevant to the current presentation and excludes clinician interpretations or derived findings. 3) The style constraints control how the content is expressed, aligning register and cadence to the persona so that voice, complexity, and tone reflect the patient's background. Full specifications appear in Figure 1.

3.3. Evaluation Framework

Rubric-based evaluation We assess each self-report with a rubric that captures both clinical fidelity and patient-style quality across four dimensions: Clinical Correctness, Faithfulness, Persona alignment, and Realism. We prioritize Clinical Correctness and Faithfulness in the overall score, assigning weights of 0.40 and 0.30; Persona alignment and Realism carry secondary weights of 0.20 and 0.10. For Clinical Correctness, we score only the claims explicitly made in the report and compare them with the patient-stated EHR; omissions of rele-

Table 1: Summary of linguistic metrics. (Avg) Length: mean tokens, sentences, and tokens/sentence. Lexical diversity (LexDiv) = unique word types ÷ total tokens; FK (Flesch–Kincaid) grade estimates the U.S. school grade level needed to read the text, higher is harder. POS / Content (%): share of content words (ADJ+ADV+VERB+NOUN) and each POS.

Generator	(Avg) Length tokens | sents | sent len	Lexicon & Readability Vocab | LexDiv% | FK	POS / Content (%) Content | Adj | Adv | Verb | Noun
Origin	$153.45 \mid 7.22 \mid 21.70$	$101.97 \mid 68.06 \mid 12.89$	$51.70 \mid 10.17 \mid 4.45 \mid 13.35 \mid 23.73$
GPT-4o-mini	$99.77 \mid 5.18 \mid 19.37$	$71.39 \mid 72.66 \mid 9.79$	$50.52 \mid 8.22 \mid 4.45 \mid 13.76 \mid 24.09$
GPT-5-mini	$103.96 \mid 3.74 \mid 30.91$	$72.75 \mid 71.40 \mid 15.08$	$53.10 \mid 10.06 \mid 4.39 \mid 13.42 \mid 25.23$

vant symptoms do not incur penalties, whereas contradictions and unsupported medical facts do. For Faithfulness, we require that all content originate from the patient-stated EHR and exclude clinician interpretations, exam or test results, diagnoses, and treatment speculation. Persona alignment and Realism evaluate the naturalness and plausibility of the narrative, capturing whether the language, tone, and structure align with how a real patient would describe their symptoms. Two independent graders, GPT-4.1 and Claude Sonnet 3.7, apply the same rubric to each case, yielding comparable scores for cross-grader validation and fairness analysis. This framework standardizes evaluation, reduces subjective bias, and keeps the assessment focused on how accurately and authentically the reports reflect patient-stated content. Detailed scoring criteria appear in Figure 2.

Rubric-free evaluation While the rubric-based framework provides transparency and consistency, it may overlook aspects of quality not explicitly defined in the scoring criteria. To capture these broader judgments, we introduce a rubric-free evaluation. In this setting, the model receives the EHR excerpt, the associated persona, and two candidate self-reports generated by different systems. It is then prompted to choose the report that better matches the EHR's patient-stated facts, sounds natural in a first-person patient voice, and fits the persona's tone. This open-ended comparison allows evaluators to express holistic preferences that may include subtle linguistic and stylistic cues beyond the rubric dimensions. To prevent position bias, we randomize the order of the two reports during evaluation.

4. Experiments

4.1. Experimental Setup

Data We draw 10000 encounters from MIMIC–IV using stratified sampling over age, gender, primary language, insurance, admission setting/type, and case complexity. Each stratum contributes a fixed quota to ensure uniform coverage and stable per-stratum estimates.

Generation We use low-variance decoding to ensure consistent and realistic first-person narratives across models. For GPT-5-mini, generation follows the model's default deterministic decoding, without explicit temperature or sampling parameters, as these are not configurable in the current Responses API. For GPT-4o-mini, we apply a low temperature of 0.2 and nucleus sampling with $p = 0.95$ to balance fluency and determinism. Both models generate a single paragraph of roughly 120 words with a maximum limit of 350 tokens. The statistics of the generated reports are presented in Table 1.

Evaluation Our experiments evaluate the report generation quality of GPT-4o-mini and GPT-5-mini through two complementary protocols: rubric-based scoring and rubric-free preference. Both evaluation settings include comprehensive significance testing and inter-grader consistency analyses to ensure the results are statistically sound and reliable.

4.2. Significance and agreement metrics

Cohen's d (standardized mean difference). A scale-free effect size that expresses a mean difference in standard-deviation units. For independent samples,

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}, \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

For paired samples,

$$d = \frac{\bar{\Delta}}{s_\Delta}, \qquad \Delta = x_1 - x_2.$$

Conventional guidelines treat $|d| \approx 0.2, 0.5, 0.8$ as small, medium, and large effects.

Cohen's g (preference effect size). A directional effect size for a binary preference. If p is the win proportion among non-ties, then

$$g = p - 0.5.$$

Table 2: Comparison of two graders—GPT–5 and Claude Sonnet 3.7—on three generators (Origin=no-persona baseline, GPT–4o-mini, GPT–5-mini). Cells show mean ± SD across cases for *Clinical Correctness*, *Faithfulness*, *Persona*, *Realism*, and *Overall* (higher is better). The last two rows in each block report the MAE between graders $\overline{|\text{diff}|}$ and the tolerance agreement $P(|\text{diff}| \leq 0.50)$.

Generator	Grader	Clinical Correctness	Faithfulness	Persona	Realism	Overall
Origin	GPT-4.1	4.23 ± 1.14	4.38 ± 0.80	4.76 ± 0.26	4.56 ± 0.31	4.39 ± 0.60
	Sonnet 3.7	4.25 ± 1.08	4.21 ± 1.21	4.01 ± 0.35	4.08 ± 0.32	4.17 ± 0.71
	$\overline{\lvert\text{diff}\rvert}$	0.359	0.375	0.765	0.615	0.370
	$P(\lvert\text{diff}\rvert \leq 0.50)$	0.769	0.808	0.473	0.631	0.799
GPT-4o-mini	GPT-4.1	4.54 ± 0.50	4.87 ± 0.12	4.95 ± 0.05	4.90 ± 0.09	4.75 ± 0.15
	Sonnet 3.7	4.67 ± 0.30	4.87 ± 0.19	4.30 ± 0.14	4.46 ± 0.15	4.63 ± 0.15
	$\overline{\lvert\text{diff}\rvert}$	0.353	0.186	0.653	0.514	0.277
	$P(\lvert\text{diff}\rvert \leq 0.50)$	0.832	0.890	0.627	0.820	0.884
GPT-5-mini	GPT-4.1	4.97 ± 0.03	4.99 ± 0.01	4.94 ± 0.05	4.75 ± 0.20	4.94 ± 0.02
	Sonnet 3.7	4.97 ± 0.04	4.98 ± 0.02	4.32 ± 0.10	4.42 ± 0.11	4.79 ± 0.03
	$\overline{\lvert\text{diff}\rvert}$	0.043	0.027	0.634	0.513	0.183
	$P(\lvert\text{diff}\rvert \leq 0.50)$	0.974	0.988	0.666	0.805	0.983

Table 3: Paired comparison of GPT-5-mini versus GPT-4o-mini across metrics and graders. The difference is $\Delta = \text{score}_{\text{5-mini}} - \text{score}_{\text{4o-mini}}$, so positive values favor GPT-5-mini.

Metric	Grader	Mean Δ	95% CI	Cohen's d	x_{90}
Clinical Correctness***	GPT-4.1	0.435	[0.421, 0.449]	0.61	1.00
	Claude 3.7	0.301	[0.290, 0.311]	0.55	1.00
Faithfulness***	GPT-4.1	0.120	[0.113, 0.127]	0.33	1.00
	Claude 3.7	0.110	[0.101, 0.118]	0.26	0.50
Persona	GPT-4.1	−0.005	[-0.010, 0.000]	−0.02	0.00
	Claude 3.7	0.022	[0.015, 0.030]	0.06	0.50
Realism***	GPT-4.1	−0.153	[-0.163, -0.143]	−0.30	1.00
	Claude 3.7	−0.033	[-0.041, -0.024]	−0.07	0.50

Notes. $\Delta > 0$ favors GPT-5-mini and $\Delta < 0$ favors GPT-4o-mini. The 95% CI reports the confidence interval for the mean difference. x_{90} is the smallest x such that $P(|\Delta| \leq x) \geq 0.90$. Significance codes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. For *Persona*, GPT-4.1 is not significant ($p \approx .062$) whereas Claude 3.7 is significant ($p < .001$).

Values $g > 0$ indicate preference for the system, $g < 0$ indicate preference against it, and $|g|$ reflects the strength of preference.

Cohen's κ (inter-rater agreement). An agreement coefficient that corrects for chance agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed agreement and P_e is the expected agreement computed from the raters' marginal distributions. Here $\kappa = 1$ denotes perfect agreement, $\kappa = 0$ corresponds to chance-level agreement, and negative values indicate systematic disagreement. Applicable to binary and multiclass settings.

Wilson confidence interval for a binomial proportion. A robust approximate interval for a proportion p based on k successes in n trials. With $z = z_{1-\alpha/2}$,

$$\hat{p}_{\text{W}} = \frac{k + \frac{z^2}{2}}{n + z^2}, \quad \text{half-width} = \frac{z}{n + z^2}\sqrt{\frac{k(n-k)}{n} + \frac{z^2}{4}},$$

so the interval is $\hat{p}_{\text{W}} \pm \text{half-width}$. This interval performs well for moderate n and for p near 0 or 1.

b_{01} and b_{10} (McNemar discordant counts). For paired binary outcomes from two raters or classifiers, b_{01} counts cases labeled negative by rater 1 and positive by rater 2, and b_{10} counts the opposite direction. These off-diagonal cells capture directional disagreement and serve as inputs to McNemar's test.

McNemar's χ^2 statistic. A test of the null hypothesis $H_0 : b_{01} = b_{10}$ that discordant outcomes are symmetric. The continuity-corrected statistic is

$$\chi^2 = \frac{(|b_{01} - b_{10}| - 1)^2}{b_{01} + b_{10}},$$

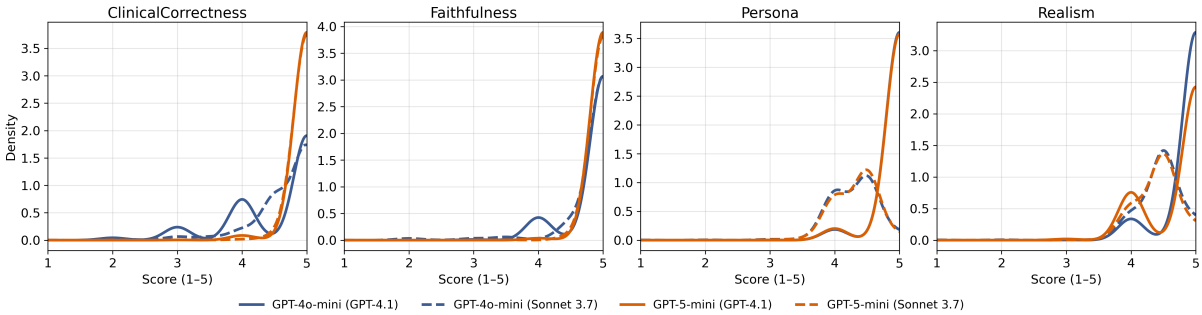

Figure 3: Score distributions (KDE) for four evaluation metrics: *ClinicalCorrectness*, *Faithfulness*, *Persona*, and *Realism*.

Table 4: Pairwise preference summary. Outcome proportions show which report (4o vs. 5m) each grader preferred. "5m win rate (excl. tie)" uses non-tie cases with Wilson 95% CI; $g = \text{winrate} - 0.5$ (Cohen's g). The last row reports inter-grader agreement (rows=GPT-4.1, cols=Claude 3.7).

Grader	4o (%)	5m (%)	tie (%)	5m win rate (excl. tie)	Cohen's g
GPT-4.1	42.58	25.18	32.24	37.66% [36.89%, 38.42%]	−0.123
Claude 3.7	23.98	62.65	13.37	72.32% [71.61%, 73.03%]	0.230
Inter-grader agreement (rows = GPT-4.1, cols = Claude 3.7):					
κ (3-class): 0.070	κ (binary): 0.063	$b_{01} = 2042$	$b_{10} = 2656$	$\chi^2 = 1442.706$	

Notes. For each grader, the "5m win rate (excl. tie)" is tested against 50% with a two-sided binomial test; both are significant ($p < .001$, omitted from table). g is Cohen's g (winrate-0.5). Inter-grader agreement reports Cohen's κ for 3 classes and binary settings (ties removed), and McNemar discordant counts b_{01} (GPT-4.1 chose 4o, Claude 3.7 chose 5m) and b_{10} (vice versa) with χ^2 statistic.

which is approximately chi-square distributed with one degree of freedom. A significant result indicates asymmetric disagreement between the two raters.

5. Results

5.1. Rubric-based Evaluation

Overall Results. Table 2 summarizes rubric-based scores across systems and graders. Performance increases progressively from the no-persona baseline to GPT-4o-mini and further to GPT-5-mini, with the largest improvements in Clinical Correctness and Faithfulness. On these factual dimensions, both graders converge in their assessments for GPT-5-mini, suggesting a shared recognition of its stronger grounding in patient-stated information. Residual variation is concentrated in stylistic dimensions—Persona and Realism—where Claude Sonnet 3.7 assigns lower ratings than GPT-4.1, indicating that differences stem primarily from stylistic expectations rather than factual accuracy. Overall, GPT-5-mini achieves clearer factual gains, while cross-grader differences mainly reflect style judgments.

Agreement and Statistical Analysis. To validate these trends, Table 3 reports inter-grader agreement and significance statistics for each rubric dimension across GPT-4.1 and Claude Sonnet 3.7. Both graders show consistent and significant improvements for GPT-5-mini in Clinical Correctness and Faithfulness (Wilcoxon $p < .001$), with positive mean differences and moderate effect sizes (Cohen's $d \approx 0.5{-}0.6$), confirming closer alignment with patient-stated facts. Effects on Persona and Realism are weaker and less consistent: GPT-4.1 finds no significant change in Persona, whereas Claude Sonnet 3.7 reports a small improvement; both graders assign slightly lower Realism to GPT-5-mini. Taken together, the results confirm robust factual gains from GPT-5-mini, while residual variation largely arises from grader-dependent stylistic preferences.

Score Distribution Analysis. Figure 3 visualizes the score distributions across all four rubric dimensions. Most distributions are compact and right-skewed, with densities concentrated above 3.5, indicating generally high evaluation scores. The main exception is Clinical Correctness for GPT-4o-mini, whose distribution spreads further left and dips below 3.5 more often than the others. Persona and Realism cluster near the upper range but remain broader than the factual dimensions, reflecting greater variability in stylistic assessments.

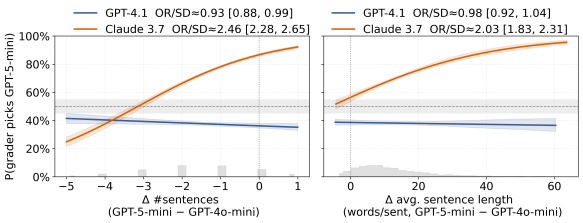

Figure 4: Sentence structure and grader preference. Left: probability of selecting *GPT-5-mini* vs the gap in number of sentences (5-mini−4o-mini). Right: same vs the gap in mean sentence length (words per sentence). Curves are binned logistic fits with 95% confidence bands; the horizontal line marks 50% and the vertical line marks $\Delta = 0$. Legends report odds ratios per one standard deviation. Claude 3.7 favors longer and denser outputs (OR/SD ≈ 2.46 for sentence count; ≈ 2.03 for length), whereas GPT-4.1 is near flat (OR/SD ≈ 0.93 and ≈ 0.98).

Overall, these patterns corroborate that GPT-5-mini improves factual consistency while stylistic judgments remain more subjective across graders.

5.2. Rubric-free Evaluation

Overall Results. To capture holistic judgments beyond the predefined rubric, we conducted a rubric-free pairwise preference test between GPT-4o-mini and GPT-5-mini. Table 4 shows that the two graders diverge sharply: GPT-4.1 prefers GPT-4o-mini, while Claude Sonnet 3.7 favors GPT-5-mini. Both deviations from a 50–50 split are statistically significant ($p < 0.001$), yet inter-grader agreement remains low ($\kappa \approx 0.07$), confirming that stylistic bias rather than content differences drives disagreement. McNemar's test on discordant cases further indicates directional asymmetry between graders.

Bias and Reliability. Figure 4 shows that grader preferences are systematically shaped by surface form: Claude Sonnet 3.7 consistently favors longer or denser reports, with preference increasing alongside the number of sentences and mean sentence length (OR/SD $\approx$ 2.0–2.5), whereas GPT-4.1 remains largely insensitive to verbosity, indicating a more content-oriented criterion. Taken together, these patterns suggest that apparent disagreement stems from structured stylistic bias rather than random inconsistency, underscoring the need to pair rubric-free judgments with rubric-based factual assessments.

6. Case Study

Table 5 reports the relationship between education level and the Flesch–Kincaid grade of the generated reports. We expect self-reports to align with the assigned persona, and education is a readily verifiable attribute that strongly shapes linguistic complexity and style, so we use it as a case study. Both generators follow the intended control: the mean Flesch–Kincaid grade increases monotonically from *primary* to *PhD*, indicating that the narratives adjust complexity to the assigned education persona. However, GPT-5-mini consistently produces higher readability grades than GPT-4o-mini at every level, reflecting more elaborate vocabulary and syntax. To balance stylistic realism and diversity, we adopt a mixed-generation strategy in which lower-education personas (primary to high school) are generated with GPT-4o-mini and higher-education personas (college to PhD) with GPT-5-mini. We will use this hybrid protocol when generating and releasing the final version of the dataset.

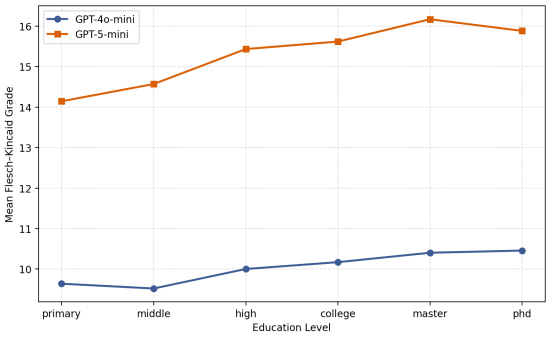

Figure 5: Mean Flesch–Kincaid grade by simulated education level. GPT-5-mini consistently produces more complex language than GPT-4o-mini, with readability rising alongside education level.

7. Conclusion

This study introduces a framework for generating realistic, persona-conditioned patient self-reports directly from structured EHR data. By conditioning large language models on social and clinical attributes such as age, gender, education, and illness severity, we reproduce how real patients describe their symptoms while maintaining strict fidelity to patient-stated facts. Through rubric-based and rubric-free evaluations, we demonstrate that GPT-5-mini achieves higher factual accuracy and faithfulness than GPT-4o-mini, whereas stylistic variation largely reflects grader-specific preferences rather than substantive differences in content. The dual-model, dual-grader design enhances interpretability and robustness in evaluation, providing a balanced foundation for dataset curation. Moving forward, we plan to release the final dataset generated under the proposed mixed-model protocol, supporting future research on patient-style text generation,

health communication, and bias-aware evaluation in clinical NLP.

8. Appendices

8.1. Ethical Considerations

All human-subject data used in this study (MIMIC-IV and MIMIC-IV-Note) are fully de-identified under HIPAA and were released under a PhysioNet Data Use Agreement (DUA), which mandates human-subjects training and strictly forbids any attempt at re-identification. The original IRBs at Beth Israel Deaconess Medical Center and MIT approved the data release with a waiver of informed consent and determined that secondary analyses of these de-identified records are exempt from ongoing review. Our use strictly adheres to these terms and remains within a research-only setting.

The generated patient self-reports are synthetic paraphrases derived from de-identified patient-stated content and contain no additional identifiable information. Access to source data is restricted to credentialed PhysioNet users, and no identifiable patient information is included in released outputs.

Persona attributes are used exclusively as stylistic controls (e.g., phrasing and narrative tone) rather than clinical or diagnostic variables. We acknowledge that persona-based conditioning may introduce or amplify unintended biases, and therefore encourage cautious downstream use and further bias-aware evaluation.

Finally, the evaluation protocol relies on LLM-based graders, which do not replace clinician judgment. While this design enables large-scale and reproducible assessment, future work will include physician-based evaluation to further validate clinical reliability.

8.2. Acknowledgements

This work was partially supported by the Data Science Center at The University of Memphis. The opinions, findings, and results are solely those of the authors.

References

Nancy D. Berkman, Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karen Crotty. 2011a. Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, 155(2):97–107.

Nancy D. Berkman, Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karin Crotty. 2011b. Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, 155(2):97–107.

John G. Canto, William J. Rogers, Robert J. Goldberg, Eric D. Peterson, Nanette K. Wenger, Viola Vaccarino, Catarina I. Kiefe, Peter D. Frederick, George Sopko, Zhiying Zheng, et al. 2012. Association of age and sex with myocardial infarction symptom presentation and in-hospital mortality. *JAMA*, 307(8):813–822.

John G. Canto, Michael G. Shlipak, William J. Rogers, Judith A. Malmgren, Catarina I. Kiefe, Harold V. Barron, Peter D. Frederick, Joseph P. Ornato, Christine M. Gibson, and William D. Weaver. 2000. Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. *JAMA*, 283(24):3223–3229.

Juan G. Canto, Emily A. Canto, and Robert J. Goldberg. 2007. Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. *JAMA*, 297(7):813–816.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: An emr-based dataset for clinical multi-step diagnosis. *arXiv e-prints*, pages arXiv–2408.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9278–9298, Abu Dhabi, UAE. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Aaron Gayles, A. Shammout, Steven Horng, Tom J. Pollard, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–10.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Stefan Lautenbacher, Jan H. Peters, Michael Heesen, Jennifer Scheel, and Miriam Kunz. 2017a. Age changes in pain perception: A systematic review and meta-analysis of age effects on pain and tolerance thresholds. *Neuroscience & Biobehavioral Reviews*, 75:104–113.

Stefan Lautenbacher, Mette L Peters, Maarten Heesen, Julia Scheel, and Miriam Kunz. 2017b. Age changes in pain perception: A systematic-review and meta-analysis of age differences in

pain thresholds and tolerance. *Neuroscience & Biobehavioral Reviews*, 75:104–113.

Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351:h3480.

Mary E Short, Ron Z Goetzel, Xue Pei, Mohannad J Tabrizi, Ronald J Ozminkowski, Teresa B Gibson, David M DeJoy, and Mark G Wilson. 2009a. How accurate are self-reports? an analysis of self-reported health care utilization and absenteeism. *Journal of Occupational and Environmental Medicine*, 51(7):786–796.

Mary Eastwood Short, Ron Z. Goetzel, Xiaofei Pei, Mohammad J. Tabrizi, and Ronald J. Ozminkowski. 2009b. How accurate are self-reports? analysis of self-reported health care utilization and absenteeism. *Journal of Occupational and Environmental Medicine*, 51(7):786–796.

Zhixiang Su, Yinan Zhang, Jiazheng Jing, Jie Xiao, and Zhiqi Shen. 2024. Enabling patient-side disease prediction via the integration of patient narratives. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 581–584, New York, NY, USA. Association for Computing Machinery.

R. E. M. van Oosterhout, A. R. de Boer, A. H. E. M. Maas, F. H. Rutten, M. L. Bots, and S. A. E. Peters. 2020a. Sex differences in symptom presentation in acute coronary syndromes: A systematic review and meta-analysis. *Journal of the American Heart Association*, 9(9):e014733.

René E. M. van Oosterhout, Ariënne R. de Boer, Angela H. E. M. Maas, et al. 2020b. Sex differences in symptom presentation in acute coronary syndromes: A systematic review and meta-analysis. *Journal of the American Heart Association*, 9(13):e014733.

Evelyn Verlinde, Nele De Laender, Stéphanie De Maesschalck, Myriam Deveugele, and Sara Willems. 2012. The social gradient in doctor–patient communication. *International Journal for Equity in Health*, 11:12.

Matthew K Wynia and Chandra Y Osborn. 2010a. Health literacy and communication quality in health care organizations. *Journal of Health Communication*, 15(Suppl 2):102–115.

Matthew K. Wynia and Chandra Y. Osborn. 2010b. Health literacy and communication quality in health care organizations. *Journal of Health Communication*, 15(Suppl 2):102–115.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020a. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Pengtao Xie, et al. 2020b. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.