CAP: A Source-Grounded Proposition Scaffold for Faithful Clinical Dialogue-to-Note Generation

Hyunkyung Lee, Jisoo Jung, Jeonguk Lee,
Jaehyo Yoo, Wooseok Han, Minkyu Kim, Gibaeg Kim
AITRICS
harrymetsally@aitrics.com

Abstract

While Large Language Models (LLMs) have advanced clinical dialogue-to-note generation, direct transcript-only prompting can still fail in clinically important ways. In real-world physician–patient encounters, clinically salient evidence is noisy, distributed across turns, and frequently revised. Transcript-only generation is therefore vulnerable to hallucinated facts, section leakage, and final-state errors. Conversely, coarse intermediate scaffolds such as basic entity extraction fail to capture negation, temporality, and evolving problem–plan relationships.

To address this, we propose **Clinical Atomic Propositions (CAPs)**, a dialogue-aware intermediate representation for LLM-based note generation. CAPs extract standalone clinical assertions while explicitly preserving essential modifiers such as verification status, temporality, speaker source, and action type. We further introduce an optional **event consolidation layer** that reorganizes CAPs into problem-oriented care bundles, explicitly linking grounded evidence to planned actions before rendering. We evaluate sectioned-note generation as the main setting and include SOAP-template rendering as an ablation. This decomposition targets faithful, controllable note generation while producing evidence-linked structure that can support future audit and verification.

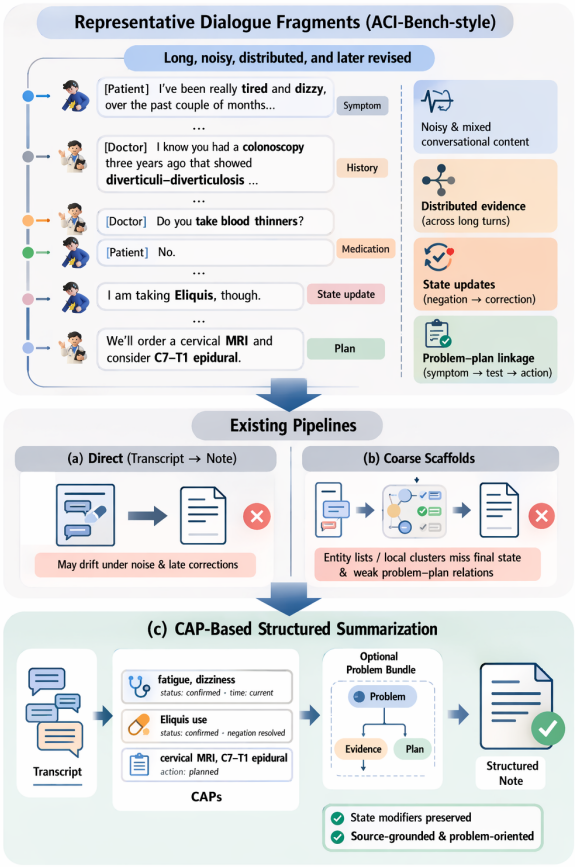

Figure 1: Why faithful dialogue-to-note generation is difficult: evidence is noisy, scattered, and later-updated; CAP makes intermediate clinical facts explicit before rendering.

1 Introduction

LLMs make it easy to prompt for a clinical note directly from a physician–patient transcript, but this convenience moves safety-critical decisions into an opaque long-context generation step (Ben Abacha et al., 2023; Giorgi et al., 2023; Krishna et al., 2021; Michalopoulos et al., 2022). The central challenge is no longer fluency alone. A useful note must be readable, *source-grounded*, clinically faithful, and section-appropriate, while preserving facts that are noisy, distributed across turns, and often revised over time. Transcript-only generation is therefore vulnerable to omission, section leakage, unsupported fill-in, and incorrect final-state tracking.

A growing line of work therefore uses *structure-first* pipelines, including section filtering, utterance clustering, and entity-centric scaffolds (Krishna et al., 2021; Michalopoulos et al., 2022; Nair et al., 2023). These approaches help by giving the final generator a smaller, more organized input. How-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 572–594
July 3-4, 2026

ever, each scaffold can lose a different clinical relation. A proximity-based cluster may split an early statement from a later correction; an entity/status list may record that a medication is present without preserving the full claim or its timing; and a section-level filter may collect relevant text without saying which problem a finding or plan supports. As Figure 1 illustrates, the failure is often not that every relevant mention is absent, but that the *final clinically grounded state* and the problem–evidence–plan structure are not preserved.

We address this gap with **Clinical Atomic Propositions (CAPs)**, a dialogue-aware intermediate representation that converts raw dialogue into standalone, source-grounded clinical assertions. CAPs preserve the modifiers that are most critical for faithful note generation, including verification status, temporality, speaker/source, and action type. A list of correct propositions is still not necessarily a note-ready structure, so we additionally introduce an optional **event consolidation layer** that groups CAPs into problem-oriented care bundles before rendering. This separates two roles that are often conflated in transcript-only generation: preserving source-grounded facts and organizing those facts into clinically coherent assessment and plan structure.

Our work makes three main contributions:

- **Source-grounded proposition scaffold.** We introduce CAPs as a proposition-level intermediate representation for clinical dialogue-to-note generation that preserves clinically important modifiers beyond entity presence alone, including verification status, temporality, speaker/source, and action type.
- **Problem-oriented structural alignment.** We design and evaluate an optional event-consolidation step, inspired by problem-oriented documentation and lightweight FHIR semantics, that groups grounded facts into problem-centered care bundles. Our analysis characterizes when this organization helps and when compression can introduce omissions.
- **Reproducible evaluation and release.** We evaluate on ACI-Bench (Yim et al., 2023) using explicit cohort filtering criteria for reproducibility. We include prompt-based reimplementations of Cluster2Sent and MEDSUM-ENT, reuse the MEDSUM-ENT GPT-R/P/F1 metric family, and adapt the same recall/precision/F1 logic to CAP-space semantic auditing. We evaluate sectioned-note rendering as the main task and SOAP-template rendering as an ablation. We release the code, prompts, postprocessing scripts, intermediate representations, and generated sectioned/SOAP notes in a public repository: `https://github.com/sallyy1/CAP-ACL-2026`.

2 Related Work

Structure-first clinical dialogue understanding and note generation. Recent work increasingly treats doctor–patient dialogue summarization as a *structure-then-generate* problem rather than pure long-context generation (Ben Abacha et al., 2023; Giorgi et al., 2023). Cluster2Sent (Krishna et al., 2021) selects section-relevant utterances and clusters nearby evidence; MedicalSum (Michalopoulos et al., 2022) and ClinicSum (Neupane et al., 2024) use guided or retrieval-filtered SOAP inputs; and MEDSUM-ENT (Nair et al., 2023) extracts entities with present/absent/unknown status labels before summarization. Adjacent work also models clinical dialogue through task-specific labels, including note-section classification and medical-order extraction (Chen et al., 2023; Corbeil et al., 2025), demonstrating that schema-constrained outputs can facilitate clinical understanding. However, existing scaffolds are usually section-level, proximity-based, task-specific, or entity/status-based. They help organize material, but do not directly represent self-contained clinical claims with source evidence, temporal status, and problem–action linkage.

From clinical IE to proposition-level representations. Clinical information extraction traditionally maps unstructured notes into entities, modifiers, and relations, and recent LLM-based IE continues this line with clinical NER/RE models and tools such as Kiwi (Hu et al., 2026). FHIR-oriented text mining and commercial systems also show the value of interoperable extraction: for example, Azure AI Language Text Analytics for health can return a FHIR resource bundle, while Azure Health Data Services provides a managed FHIR service (Daumke et al., 2019; Microsoft, 2025, 2026). These efforts support interoperability, but a dialogue-aware note scaffold must also decide which evolving claim is finally supported, how negation and temporality apply, and how evidence/actions attach to problems. Proposition-level verification work such as VeriFact(Chung et al.,

2025) motivates this granularity by decomposing clinical text into simple statements for fact checking. CAP adapts this idea before rendering: it extracts evidence-linked, FHIR-inspired propositions from dialogue, then optionally consolidates them into problem-oriented bundles for note generation.

3 Motivation and Problem Setting

Clinical dialogue-to-note generation is difficult because clinically important evidence is noisy, distributed, and frequently revised across the encounter. This creates four recurring challenges for transcript-only or coarse-scaffold generation.

Motivation challenges.

- **Noisy dialogue:** salient evidence is interleaved with low-value conversational content.
- **Lay expression vs. over-abstraction:** patient language must be normalized without losing negation, uncertainty, temporality, or medication nuance.
- **Brittle state tracking:** key facts are revised over time, including symptom status, resolved uncertainty, medication use, and downstream plans.
- **Baseline limitations:** Direct is vulnerable to noise and section leakage, Cluster2Sent is local, and MEDSUM-ENT can miss problem-linked organization and state–plan separation.

Given a transcript X, our goal is to generate a clinical note Y that is clinically faithful, section-appropriate, and source-grounded. We ask:

Research questions.

- **RQ1: Source-grounded preservation.** Can a normalized intermediate representation better preserve clinically salient source content?
- **RQ2: Clinical state fidelity.** Can it better retain negation, uncertainty, temporality, and medication state or change?
- **RQ3: Problem-oriented organization.** Can it better support problem–evidence–plan linkage and final note structure?

Figure 2 illustrates the CAP pipeline, while Tables 1, 2, and 3 summarize the running example, FHIR-inspired CAP-type schema, and compared method scaffolds.

4 Method

4.1 Pipeline Overview

We decompose dialogue-to-note generation into a structure-first pipeline:

$$X \rightarrow C \rightarrow E \rightarrow Y, \tag{1}$$

where X is the dialogue transcript, C is a set of **Clinical Atomic Propositions (CAPs)**, E is an optional **problem-oriented consolidation plan** ("event" bundles), and Y is the rendered clinical note. CAPs provide a source-grounded content inventory, while event consolidation aligns that inventory to problem-oriented documentation practice (POMR/SOAP) by grouping evidence and actions under an anchor problem. The renderer then writes the final sectioned or SOAP note from the selected scaffold.

4.2 Clinical Atomic Propositions

A **Clinical Atomic Proposition (CAP)** is a standalone clinical assertion that is verifiable against the dialogue and close to the semantic grain of a note sentence. A CAP is not just an entity mention. It records the clinical claim plus state modifiers such as verification status, temporality, speaker/source, numeric attributes, and whether the claim describes current state or an intended action. CAPs are extracted automatically by prompted, schema-constrained JSON generation; we do not manually author them at test time.

The CAP schema is *FHIR-inspired* rather than FHIR-native (Health Level Seven International, 2019; Ayaz et al., 2021). Table 2 lists the canonical `cap_types` used in this paper. While FHIR is designed primarily for interoperable health-data exchange, CAP adopts its resource-oriented separation of clinical concepts to define a compact middle schema for note generation. This schema separates condition-like problems, observation-like evidence, current medication state, medication changes, and planned actions. The distinction is central for dialogue: for example, "taking Eliquis" and "order an MRI" should not be collapsed into the same kind of fact. The small CAP-extraction dev check is reported in Appendix B.

4.3 Event Consolidation and Rendering

Consistent with the Problem-Oriented Medical Record (POMR) (Weed, 1968), clinical notes are typically organized around problems with linked

Stage	D2N026-style late medication update
Dialogue X	"Do you take blood thinners?" → "No." → later correction: "I am taking **Eliquis**, though." The physician also discusses a **cervical MRI** and tentative **C7–T1 epidural**.
CAPs C	Separate source-grounded propositions preserve the corrected final state, e.g., `MedicationStatement: patient currently taking Eliquis; Order: cervical MRI planned; ProcedurePlan: C7-T1 epidural discussed`.
Bundle E	The bundler groups these CAPs under an anchor problem such as `Problem: cervical radiculopathy`, with linked evidence and plan slots.
Note Y	The renderer realizes the corrected state and plan in prose, e.g., Eliquis use, cervical MRI, and possible C7–T1 epidural.

Table 1: Running example for the proposed $X \rightarrow C \rightarrow E \rightarrow Y$ pipeline.

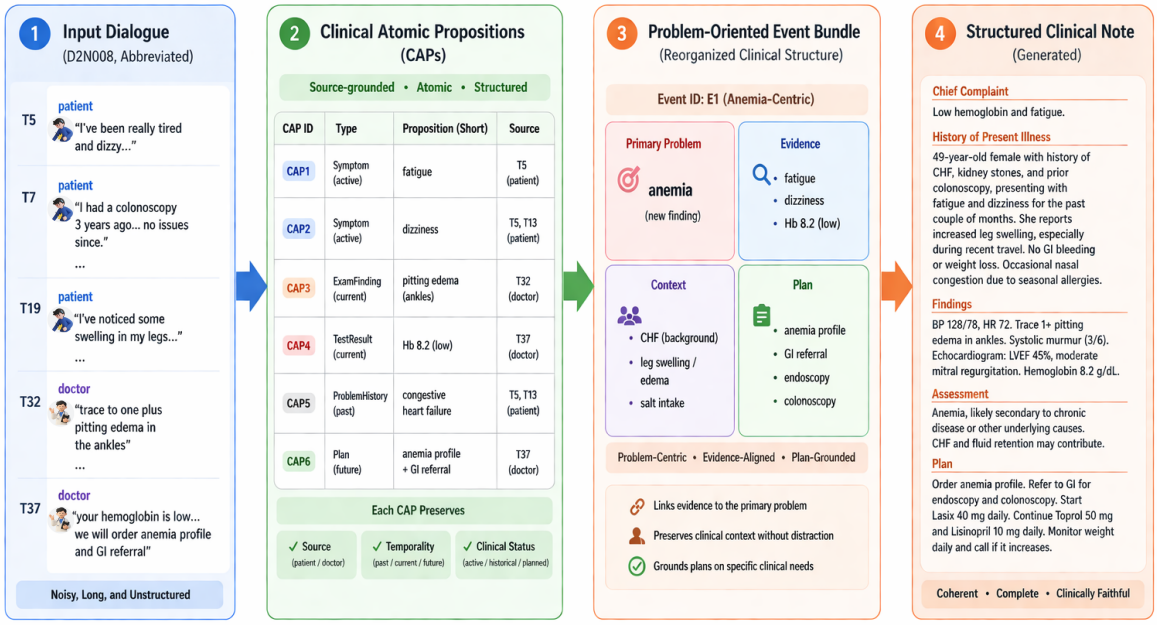

Figure 2: Overview of the CAP-based pipeline. Dialogue is first converted into source-grounded atomic propositions (CAPs), which are then reorganized into a problem-oriented event structure. This structured representation enables coherent and clinically faithful note generation.

evidence and plans, a structure commonly operationalized through the SOAP (Buchanan, 2017) template. Our optional event layer turns a flat CAP list into **problem-oriented bundles**: each bundle anchors on a problem and links related subjective evidence, objective evidence, and plan actions. We treat this as an alignment module rather than a guaranteed quality booster. It can make problem–plan linkage explicit, but it can also over-compress or omit facts if the bundling is too aggressive.

The final renderer receives the chosen scaffold and writes a sectioned or SOAP note. In the main CAP and CAP+Event settings, the transcript is also available for surface realization; transcript-free ablations in Appendix E test how much the scaffold can support rendering on its own. Prompt templates and baseline reimplementation details are provided in Appendix C.

5 Experiments

Dataset. We use ACI-Bench (Yim et al., 2023), a publicly documented ambient clinical intelligence benchmark for automatic visit-note generation from doctor–patient conversations. The original paper describes ACI-Bench as a *shared open dataset* that is *freely available*; we follow that release as our primary data source. Following the benchmark setup, we use the provided conversation transcripts as model input and the provided reference notes as evaluation targets.

To avoid ambiguity about supervision, we do not create new gold notes or re-annotate the benchmark. Reference-note provenance and annotation protocol are therefore inherited from ACI-Bench and are documented in Yim et al. (2023). For reproducibility, we start from 207 benchmark encounters

CAP type	Intended semantics	FHIR-inspired analogue	Key state fields
`ChiefComplaint`	Reason for visit / presenting concern	Encounter reason (conceptual)	`verification_status`, `temporality`
`Problem`	Active problem / diagnosis / clinical concern	Condition-like	`verification_status`, `clinical_status`, `temporality`
`ProblemHistory`	Historical problem context	Condition-like (historical)	`clinical_status=historical`
`ExamFinding`	Physical exam finding / symptom-style observation	Observation-like	`verification_status` (e.g., negated), `temporality`
`TestResult`	Lab / imaging result	Observation/Report-like	`verification_status`, numeric `attributes`
`MedicationStatement`	Current medication state (what the patient is taking)	MedicationStatement-like	`clinical_status=active` vs. historical
`MedicationRequest`	Intended medication start/stop/change	MedicationRequest-like	`clinical_status=planned`
`Order`	Concrete order (labs, imaging, referral)	ServiceRequest-like	`clinical_status=planned`
`FollowUp`	Follow-up timing / appointment plan	CarePlan/Appointment-like	`temporality`, `clinical_status=planned`
`Counseling`	Advice / behavior counseling / education	CarePlan-activity-like	`temporality`, `clinical_status`

Table 2: Canonical `cap_types` and intended semantics. The mapping is a lightweight FHIR-inspired abstraction, not a full FHIR implementation. CAPs additionally store `proposition_text`, `canonical_concept`, `evidence`, and optional `linked_problem` pointers used by event consolidation.

Method	Input to downstream renderer	Intermediate scaffold	Unit	Value / limitation
Direct (base)	Transcript X	None	None	+ fluent baseline – noise, leakage, state drift
MEDSUM-ENT (prior)	Transcript X + entity/status scaffold	6-category plan (*Demographics/SDOH*, *Intent*, *Positives*, *Negatives*, *Unknowns*, *History*) $\times$ {*present*, *absent*, *unknown*}	Entity/status	+ status cues – weak problem–plan linkage
Cluster2Sent (prior)	Section-targeted evidence units only (no X)	Utterance clusters (proximity; τ=2)	Utterance cluster	+ local coherence – long-range/state ambiguity
CAP (ours)	Transcript X + CAP set C	Atomic propositions with clinical modifiers	Proposition	+ grounded state fidelity – needs rendering for organization
CAP+Event (ours)	Transcript X + C + bundle plan E	Problem-oriented bundles linking evidence and actions	Problem bundle	+ problem–plan organization – compression can omit

Table 3: Strategic positioning of the five main compared methods and their intermediate scaffolds. Method-label shading matches Table 6: gray marks the Direct baseline, blue marks transcript+entity/proposition scaffold methods, and beige marks clustering/bundling scaffold methods. Transcript-free ablations are reported in Appendix E.

and apply a predefined CAP-quality cohort filter, yielding a final 197-case cohort (Appendix A lists excluded IDs and exact criteria). Our release package focuses on reproducible prompts, intermediate representations, and scoring scripts for this fixed cohort.

Setup and methods. We evaluate on the final 197-case ACI-Bench cohort after CAP extraction and predefined quality filtering; cohort details and excluded IDs are in Appendix A. The main experiments use a sectioned-note template, with SOAP ablations in Appendix D. We compare

Aspect	GPT-R/P/F1 (Nair-style)	semCAP-R/P/F1 (ours)
Primary purpose	Reference-oriented overlap check	Source-grounded proposition audit
Compared units	Medical concepts with status cues; extracted under MEDSUM-ENT-style categories (*Demographics/SDOH, Intent, Positives, Negatives, Unknowns, History*)	Proposition-level CAP units aligned to dialogue evidence
Source of truth	Reference note (Generated note ↔ Reference note)	Dialogue-grounded CAP/audited propositions (Dialogue transcript ↔ Generated note)
Rewards	Coverage of reference-like concepts and *Present/Absent/Unknown* status consistency	Preservation of grounded assertions and key modifiers (*verification, temporality, source, action type*, etc.)
May miss	Direct dialogue grounding and problem–plan linkage even when concept overlap is high	Broader narrative/note-quality aspects
Role in this paper	Complementary note-level metric	Primary targeted metric for CAP claim
Limitation	Reference-oriented, not grounding-specific	May be partially aligned with CAP representation

Table 4: Comparison of Nair-style GPT metrics and semCAP. We report both because they capture complementary aspects: reference-note overlap vs. dialogue-grounded faithfulness.

View (D2N026)	Concrete extracted units (abridged)	Count / score
Transcript CAP units (semCAP-R)	"Left arm/hand pain for about two weeks"; "Hand weakness and grip difficulty"; "Patient is taking Eliquis"; "Clinician orders cervical MRI", ...	25 CAPs
Generated-note CAP units (semCAP-P, CAP note)	"The patient reports left arm pain."; "...hand pain began approximately two weeks ago."; "I will order a cervical MRI."; "...tentatively schedule a left C7–T1 epidural.", ...	21 CAPs
Reference-note concepts (GPT-R)	*left arm pain*, *gabapentin*, *cervical MRI*, *left C7–T1 epidural*, *blood thinners*, ...	45 concepts
Generated-note concepts (GPT-P, CAP note)	*left arm pain*, *hand weakness*, *cervical radiculopathy*, *cervical MRI*, *C7–T1 epidural*, ...	24 concepts
Case-level metric snapshot (sectioned template)		
Direct	Nair GPT-F1 / semCAP-F1	0.098 / 0.000
CAP	Nair GPT-F1 / semCAP-F1	0.488 / 0.543

Table 5: Concrete example of the two metric families on D2N026 (values from released case-level outputs). Nair-style GPT metrics compare reference-note vs generated-note concepts, while semCAP compares transcript-grounded vs generated-note CAP propositions. The Direct semCAP-F1 value of 0.000 is a case-level boundary example under strict proposition matching, not a global trend. Full R/P/F1 values are reported in Table 6.

Direct, **Cluster2Sent** (Krishna et al., 2021), **MEDSUM-ENT** (Nair et al., 2023), **CAP**, and **CAP+Event**; Table 3 summarizes their inputs, scaffolds, and limitations. All methods use the same downstream generator, and all CAP extraction/rendering stages use the open-weight **Gemma-3-4B-Instruct** model (Hugging Face ID: `ISTA-DASLab/gemma-3-4b-it-GPTQ-4b-128g`) to reflect privacy-constrained clinical deployment. Baseline details and prompts are in Appendix C; transcript-free ablations and cost analyses are in Appendices E and F.

Metrics. We separate two automatic metric families. **GPT-R/P/F1 (Nair)** follows the MEDSUM-ENT GPT-style concept matching framework (Nair et al., 2023). Here, a *concept* denotes a medically relevant item (e.g., symptom, condition, medication, test/procedure, status cue). The framework uses a two-stage LLM process: a concept extractor first identifies candidate concepts in generated/reference notes, then a verifier judges concept matches across notes under relaxed semantic equivalence, explicitly allowing clinically valid rephrasings. GPT-recall measures reference concepts recovered by the generated note, GPT-precision measures generated concepts supported by the reference note, and GPT-F1 is their harmonic mean. This family is therefore *reference-oriented* and primarily reflects concept-level note overlap.

semCAP-R/P/F1 (ours) is our proposition-grounded semantic audit, measuring whether transcript-derived CAP propositions are preserved

Method	semCAP-R (ours)↑	semCAP-P (ours)↑	semCAP-F1 (ours)↑	GPT-R (Nair)↑	GPT-P (Nair)↑	GPT-F1 (Nair)↑
Direct (base)	0.574	0.175	0.211	0.638	0.808	**0.703**
MEDSUM-ENT (prior)	0.589	0.168	0.205	**0.640**	0.761	0.687
Cluster2Sent (prior)	0.522	0.178	0.204	0.587	**0.813**	0.672
CAP (ours)	**0.694**	**0.212**	**0.256**	0.632	0.795	0.695
CAP+Event (ours)	0.570	0.181	0.214	0.578	0.773	0.649

Table 6: Results on the 197-case ACI-Bench cohort. (Sectioned-template) Method-label shading groups the Direct baseline (gray), transcript+entity/proposition scaffold methods (blue), and clustering/bundling scaffold methods (beige). semCAP metrics are our source-grounded proposition audit; GPT metrics follow the MEDSUM-ENT concept-matching family (Nair et al., 2023). Broader checklist and PDSQI scores are reported in Appendix D.

Method vs. Direct	**semCAP-R** W/T/L, Δ	**semCAP-F1** W/T/L, Δ	**GPT-F1 (Nair)** W/T/L, Δ
MEDSUM-ENT	111/13/73, +0.015	55/99/43, -0.006	101/4/92, -0.016
Cluster2Sent	60/15/122, -0.052	47/97/53, -0.007	64/2/131, -0.031
CAP (ours)	**170/4/23, +0.120**	**83/97/17, +0.046**	85/2/110, -0.008
CAP+Event (ours)	103/15/79, -0.004	53/97/47, +0.003	83/2/112, -0.054

Table 7: Case-level win/tie/loss against Direct. (Sectioned-template) Full counts corresponding to Figure 3.

and supported in the generated note, including clinically critical modifiers (negation/state, temporality, and source grounding). semCAP is thus targeted to source-grounded faithfulness rather than broad note-quality similarity. Because semCAP is aligned with our representation, some metric-method alignment bias may remain; we therefore report semCAP jointly with the reference-oriented GPT metrics and interpret them as complementary, not interchangeable, signals. All automated metric evaluation uses **OpenAI GPT-5.1** (API model ID: `gpt-5.1-2025-11-13`) as judge. Additional checklist/PDSQI analyses are in Appendix D; Tables 4–5 summarize metric differences and one case-level example.

Main results on the 197-case cohort. Table 6 shows that CAP is strongest on all source-grounded semCAP metrics, improving semCAP-R from 0.574 for Direct to 0.694 (+0.120 absolute) and semCAP-F1 from 0.211 to 0.256. On the MEDSUM-ENT GPT concept metrics, CAP remains close to Direct (GPT-F1 0.695 vs. 0.703), while MEDSUM-ENT narrowly leads on GPT-R and Cluster2Sent on GPT-P. We therefore do not claim uniform superiority on generic note/reference overlap. Instead, the main quantitative finding is more specific: CAP improves preservation of transcript-grounded clinical propositions while maintaining competitive concept-level note similarity.

Figure 3 further clarifies the aggregate pattern. CAP wins over Direct on semCAP-R in 86.3% of cases and loses in only 11.7%, while its GPT-F1 is nearly tied on average. This supports our intended positioning: CAP is not primarily a fluency or broad quality optimizer, but a source-grounded content scaffold for preserving clinically salient facts. CAP+Event does not uniformly outperform CAP, reinforcing our view that event consolidation is an alignment module whose benefit is case dependent. Detailed win/tie/loss counts for semCAP-R/F1 and GPT-F1 are reported in Table 7.

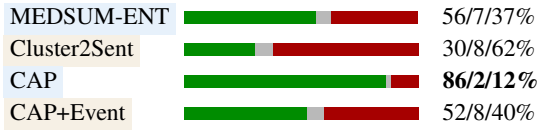

Figure 3: Case-level semCAP-R win/tie/loss rates against Direct. (Sectioned-template) Label shading matches Table 6: blue marks transcript+scaffold methods and beige marks clustering/bundling methods. Green/gray/red bars indicate wins/ties/losses.

Category-level GPT submetrics were also inspected during development, but they were sparse and unstable on this cohort. We therefore report them only as internal diagnostics and keep the main claims focused on semCAP and GPT aggregate metrics.

Qualitative evidence. D2N008 is our main qualitative case because it mixes a new anemia workup with background CHF, edema, prior GI history,

Method	Representative output snippet
Direct	Clinically plausible but structurally shallow: anemia is identified and the GI workup is mentioned, yet the CHF–anemia relation remains vague and the note never fully reconstructs anemia as the dominant organizing problem.
Cluster2Sent	Preserves fatigue and dizziness, but also keeps lower-priority content such as [incorrect emphasis] "mild nasal congestion related to seasonal allergies", producing *salience dilution* rather than a clear anemia-centered hierarchy.
MEDSUM-ENT	Entity-centric planning captures concept/status cues but does not explicitly reconstruct anemia-centered problem–evidence–plan structure; in this case the sectioned renderer can under-generate to an *empty skeleton* (see Appendix H).
CAP	Stronger structure and symptom/history coverage, but it introduces [incorrect attribution] "CHF, diverticulosis, and seasonal allergies are contributing factors", which over-expands peripheral findings into the anemia assessment.
CAP+Event (ours)	Re-centers the note on anemia as the primary problem and best preserves problem–plan linkage, including [correct] "ordering an anemia profile" and [correct] "refer ... for an endoscopy and repeat colonoscopy".

Table 8: Main qualitative case study on D2N008. Under a multi-condition scenario with background CHF and a new anemia workup, CAP-based structured generation better preserves clinically relevant problem–evidence–plan linkage than transcript-only, utterance-clustering, and entity-centric baselines (more details in Appendix H).

Case	Scenario	Observed CAP+Event behavior	ΔsemCAP-R	ΔGPT-F1 (Nair)
D2N207	Multi-condition/cardiorenal visit with CKD context and possible gout flare	Produces two bundles (*chronic kidney disease*; *possible gout flare*), improving organization and grounded recall over CAP.	$+0.218$	$+0.235$
D2N153	Localized musculoskeletal injury follow-up	Collapses to one narrow *toe injury* bundle and under-generates an empty sectioned skeleton, losing much of CAP's preserved content.	-0.714	-0.771

Table 9: Representative boundary cases for event consolidation (CAP+Event minus CAP). We report one source-grounded metric (semCAP-R) and one MEDSUM-ENT-style concept metric (GPT-F1) to show the preservation/compression trade-off under two complementary metric families.

and low-value allergy content. The key challenge is whether the note reconstructs anemia as the organizing problem and keeps the anemia profile and GI referral linked to that problem.

Table 8 illustrates the pattern. Direct remains plausible but weakly organized; Cluster2Sent preserves local evidence but over-emphasizes peripheral allergy content; MEDSUM-ENT captures entity/status cues but does not explicitly reconstruct anemia-centered problem–evidence–plan structure. CAP provides a better source-grounded evidence inventory, and CAP+Event most clearly links the anemia problem to the anemia profile and GI workup. Full qualitative cases for D2N008, D2N026, and D2N024 are shown in Appendices H, I, and J, respectively; D2N026 highlights late medication-state correction, while D2N024 is a mixed boundary case.

When does event consolidation help? Event consolidation is case dependent. In aggregate, CAP+Event underperforms CAP on semCAP-R, suggesting that bundling can introduce omission through compression. However, stratifying by the number of bundles produced by the consolidator reveals a more nuanced pattern: when the consolidator produces a modest multi-bundle plan (E with two bundles), organization-oriented scores improve slightly on average (mean ΔPDSQI $+0.04$; mean ΔF-CL(A+B+C) $+0.03$), whereas single-bundle outputs show consistent degradation across grounding and quality metrics (Appendix G). Table 9 shows this boundary behavior in representative cases: consolidation helps when distinct threads are preserved, but can fail when content is over-collapsed into a narrow anchor.

Clinician deep review. Two clinicians with medical training and experience reviewing clinical notes independently performed a focused deep review on three cases representing clear gain, mixed trade-off, and boundary behavior. Together with the reviewers, we designed a checklist aligned with our Motivation challenges and Research Questions to directly probe three target axes: (A) content preservation, (B) clinical state fidelity, and (C)

problem-oriented organization. Reviewers rated eight items on a 1–5 Likert scale (1=poor, 5=excellent) and flagged clinically meaningful omission/hallucination/major errors (yes/no). An optional overall usability rating was collected for context but excluded from our main faithfulness aggregate. Full rubric details, scores, and anonymized comments are in Appendix K (Figure 5).

6 Discussion

Our framework separates two sources of difficulty in clinical note generation: *factual recovery* and *document organization*. CAPs preserve clinically meaningful atomic facts from noisy dialogue, while event consolidation reorganizes grounded facts into problem-level bundles. This separation improves controllability and helps localize failures to extraction, consolidation, or rendering. It also clarifies why semCAP and GPT metrics need not move identically: the former audits transcript-grounded proposition fidelity, whereas the latter measures reference-note concept overlap.

CAP extraction also exposes a trade-off that is often hidden in transcript-only generation: *what to preserve* versus *what to suppress*. More aggressive filtering reduces small talk and background chatter, but can under-extract evidence needed later for bundling and planning. Our extractor development results show this precision–recall tension while preserving key modifiers well on matched propositions, especially verification status (Appendix B). Future work should add explicit salience/priority modeling, including severity and clinical consequentiality, so noise suppression does not become brittle omission.

More broadly, fluency alone is not sufficient: a natural-sounding note can still be clinically misleading. CAPs provide an evidence-linked structure that may also support real-time clinician correction or post-generation audit/verification, though we leave those uses to future work. Remaining CAP+Event errors show that faithful rendering from structured clinical scaffolds is still an open biomedical NLP problem, particularly when organization requires compressing heterogeneous dialogue evidence without losing clinically relevant detail.

7 Conclusion

We presented CAP, a structure-first framework for clinical dialogue-to-note generation. CAP introduces a source-grounded proposition layer that preserves clinically meaningful assertions from noisy dialogue, and an optional event consolidation layer that organizes these assertions into problem-oriented care bundles. Together, these components provide a more controlled scaffold for note generation than transcript-only or coarse structured approaches, with the clearest gains appearing in source-grounded faithfulness, clinical state preservation, and problem-oriented note construction.

Limitations

Our work has several limitations. semCAP is a targeted, representation-aligned audit metric rather than a standalone measure of overall clinical note quality, and our experiments are limited to a single benchmark, leaving broader generalization across specialties, note styles, and clinical settings for future work. The event layer is inspired by clinical documentation practice and lightweight FHIR semantics (Health Level Seven International, 2019; Ayaz et al., 2021), but is not a formal interoperability standard. Practically, the multi-stage prompted pipeline can propagate errors from CAP extraction through event clustering to rendering, and it adds non-trivial latency and token overhead relative to Direct. Finally, the renderer can still over-abstract, introduce unsupported content, or under-generate an *empty skeleton* note in a small fraction of cases; we report these failures rather than filtering them out, and a key next step is human-in-the-loop correction at the CAP stage before final rendering.

References

Muhammad Ayaz, Muhammad F. Pasha, Mohammed Y. Alzahrani, Rahmat Budiarto, and Deris Stiawan. 2021. The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7):e21929.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Joel Buchanan. 2017. Accelerating the benefits of the problem oriented medical record. *Applied Clinical Informatics*, 8(1):180–190.

Zhuohao Chen, Jangwon Kim, Yang Liu, and Shrikanth Narayanan. 2023. Clinical note section classification

on doctor-patient conversations in low-resourced settings. In *Proceedings of the Third Workshop on NLP for Medical Conversations*, pages 1–12.

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, and 10 others. 2025. VeriFact: Verifying facts in LLM-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Jean-Philippe Corbeil, Asma Ben Abacha, Jerome Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and Francois Beaulieu. 2025. Overview of the MEDIQA-OE 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Philipp Daumke, Kai U. Heitmann, Simone Heckmann, Catalina Martínez-Costa, and Stefan Schulz. 2019. Clinical text mining on FHIR. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, volume 264, pages 83–87.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. WangLab at MEDIQA-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334.

Health Level Seven International. 2019. FHIR release 4 (R4). `https://www.hl7.org/fhir/r4/`. Accessed 2026-04-16.

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K. Keloth, Vincent J. Zhang, Ruey-Ling Weng, Cathy Shyr, Qingyu Chen, Xiaoqian Jiang, Kirk E. Roberts, and Hua Xu. 2026. Information extraction from clinical notes: Are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, 33(3):553–562.

Kundan Krishna, Yixuan Zhang, Byron C. Wallace, and Mohit Iyyer. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.

George Michalopoulos, Kyle B. Mahowald, Thomas Lin, Hyo-Jung Lee, Ke Wang, Steven Bethard, and Meliha Yetisgen. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Microsoft. 2025. Utilizing fast healthcare interoperability resources (FHIR) structuring in text analytics for health. `https://learn.microsoft.com/en-us/azure/ai-services/language-service/text-analytics-for-health/concepts/fhir`. Last updated 2025-12-05; Accessed 2026-05-19.

Microsoft. 2026. What is the FHIR service in azure health data services? `https://learn.microsoft.com/en-us/azure/healthcare-apis/fhir/overview`. Last updated 2026-05-04; Accessed 2026-05-19.

Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217.

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

Lawrence L. Weed. 1968. Medical records that guide and teach. *New England Journal of Medicine*, 278(12):652–657.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *arXiv preprint arXiv:2306.02022*.

A Cohort Selection and Evaluation Protocol

Implementation details. All CAP extraction and note rendering stages used the same open-weight generator, Gemma-3-4B-Instruct (Hugging Face model ID: `ISTA-DASLab/gemma-3-4b-it-GPTQ-4b-128g`). We served the model with tensor parallel size 1, maximum context length 32,768, GPU memory utilization 0.9, and prefix caching enabled. In our runs, inference was executed on a single NVIDIA RTX Ada 6000 or RTX A6000 GPU. We set `max-num-seqs` to 32 and used a local Hugging Face cache under the workspace directory.

Evaluation model. All automated metric evaluation used **OpenAI GPT-5.1** (API model ID: `gpt-5.1-2025-11-13`) for concept-level matching, semantic CAP auditing, and checklist scoring, while the generator was fixed across methods; these API calls were used only for offline evaluation on de-identified benchmark text, not for deployment-time generation.

We began from 207 ACI-Bench encounters and removed four parse-failure cases from stage-C extraction (D2N022, D2N059, D2N088, D2N156), leaving 203 cases. We then applied a predefined CAP-quality filter and excluded six additional cases with `transcript_cap_count`=0 or `reference_cap_count`≤ 1 (D2N010, D2N053, D2N085, D2N112, D2N126, D2N187), yielding the final 197-case cohort for aggregate metrics; D2N010 is retained only as a qualitative boundary example. Two SOAP generations were empty (D2N031 for MEDSUM-ENT; D2N181 for Cluster2Sent), and we keep them as observed system behavior rather than filtering them out.

Empty structured outputs. Because the renderer must return a structured JSON object for SOAP/sectioned templates and may emit empty strings for unsupported sections, a failure mode is an *empty skeleton* note (all keys present, section content empty), which typically appears when long or heterogeneous scaffolds are hard to map under strict format constraints. On the final 197-case cohort, this occurred most frequently for MEDSUM-ENT (9/197 in sectioned; 12/197 in SOAP; 17 unique cases across templates) and more rarely for other method/template pairs ($\leq 5/197$ each). We treat these as observed generation failures and keep them in evaluation rather than filtering them out.

B CAP Extraction Development Benchmark

To reduce the risk of silently propagating extraction errors through the modular pipeline, we performed a small internal dev evaluation focused on stage C (CAP extraction) in $X \rightarrow C \rightarrow E \rightarrow Y$. We used two clinician-labeled seed CAP sets (D2N008, D2N026) and a small derived "silver" set constructed under the same rubric for prompt-iteration diagnostics (not final benchmark claims). We score extraction with relaxed semantic concept matching (threshold 0.35), reporting concept-level precision/recall/F1 and matched-pair modifier accuracy. Table 10 summarizes the final extractor used in this paper and the core precision–recall trade-off.

C Prompt Templates and Baseline Reimplementation

To facilitate reproducibility, we summarize the prompt templates and intermediate scaffold designs used for the main compared methods. Our implementation follows the structure-first intent of prior work while keeping the downstream renderer fixed so that differences primarily reflect scaffold quality.

Cluster2Sent-inspired scaffold. Following Krishna et al. (2021), we (i) select section-targeted evidence units from the dialogue, (ii) deterministically cluster nearby evidence by turn proximity (we use τ=2), and (iii) prompt the renderer to generate one concise sentence (or bullet) per cluster and place it in the appropriate note section. The crucial constraint is that the renderer is asked to use only the extracted evidence units as source material, which reduces transcript noise but can miss long-range dependencies.

```
[Section: History of Present Illness] [Cluster 1]
Left arm/hand pain started about two weeks ago;
hand weakness and grip difficulty.
[Section: Findings] [Cluster 2] Cervical source
is suspected based on exam context.
[Section: Plan] [Cluster 3] Order cervical MRI;
tentative left C7-T1 epidural.
[Section: Medications] [Cluster 4] Later turn:
patient reports taking Eliquis.
```

MEDSUM-ENT-inspired scaffold. Following Nair et al. (2023), we extract clinically relevant concepts with status cues (*present*/*absent*/*unknown*) and organize them into six planning categories (*Demographics/SDOH*, *Patient Intent*, *Pertinent Positives*, *Pertinent Negatives*, *Pertinent Unknowns*, *Medical History*). We apply an *unknown resolver*

Dev set	n	Pred	Gold	P	R	F1	VerifAcc	ClinStateAcc	TempAcc	Unsupported
Gold seeds (clinicians)	2	24.0	35.0	0.546	0.369	0.439	0.888	0.843	0.685	11.0
Silver dev (derived)	19	24.4	48.6	0.425	0.207	0.268	0.659	0.837	0.769	14.2

Table 10: CAP extraction dev benchmark for stage C. Pred/Gold/Unsupported are per-case averages. P/R/F1 are concept-level match scores under relaxed semantic matching (threshold 0.35). Modifier accuracies (verification status, clinical status, temporality) are computed on matched CAP pairs. The derived silver set is used only for prompt-iteration diagnostics.

that removes an earlier *unknown* mention when the same concept is later resolved as *present* or *absent*. This reimplementation preserves the original multi-stage spirit of MEDSUM-ENT: prompted structured extraction, lightweight normalization/postprocessing, and a separate rendering step.

```
[Category: Pertinent Positives] [Status: Present]
The patient reports fatigue for the past two months.
[Category: Pertinent Negatives] [Status: Absent]
The patient denies blood in stool.
[Category: Pertinent Unknowns] [Status: Unknown]
The patient is unsure about recent weight change.
```

MEDSUM-ENT GPT metrics and our semCAP adaptation. For note-level automated evaluation, we follow Nair et al. (2023) in using a GPT-based matching framework that computes **GPT-recall**, **GPT-precision**, and **GPT-F1**. Operationally, this framework uses two LLM roles: (i) an extractor that identifies medically relevant *concepts* from each note (generated/reference), where concepts are item-level clinical units (e.g., symptoms, diagnoses, medications, tests/procedures, and associated status cues), and (ii) a verifier that determines cross-note matches under relaxed semantic equivalence. This design allows clinically equivalent rephrasings beyond strict string overlap. GPT-R is the fraction of reference concepts recovered by the generated note, GPT-P is the fraction of generated concepts supported by the reference, and GPT-F1 summarizes both.

Our contribution is to adapt this evaluation logic from entity/category space into proposition space. Concretely, we replace concept-level units with CAP propositions and score whether transcript-grounded propositions are preserved in the generated note. We therefore report **semCAP-R**, **semCAP-P**, and **semCAP-F1** as CAP-space counterparts. The semantic audit further allows partial support and contradiction labels, making it more sensitive to source-grounded faithfulness, final-state errors, and state-update failures than pure concept overlap. At the same time, semCAP is representation-aligned by design, so we report it together with Nair-style GPT metrics as a complementary pair: semCAP for dialogue-grounded proposition preservation, GPT metrics for reference-note concept overlap.

CAP extraction and event planning. CAP extraction uses a dialogue-aware JSON schema that enforces atomic, verifiable propositions and preserves clinically important modifiers (negation, temporality, laterality, numeric values, severity, and plan intent). Event planning groups CAPs into problem-oriented bundles with linked evidence and actions, designed to improve renderability while keeping source grounding explicit.

```
CAP (Problem): The patient reports left hand weakness for two weeks.
CAP (MedicationStatement): The patient is taking Eliquis.
CAP (Order): The clinician orders a cervical MRI.
CAP (Plan): The clinician tentatively plans a left C7–T1 epidural.
```

Code and prompts. The full implementation, including prompts for scaffold extraction, rendering, and evaluation, is publicly available together with the postprocessing rules and scripts needed to reproduce our results. We also release intermediate artifacts—extracted scaffolds, CAPs, event plans, and generated sectioned/SOAP notes—for the final filtered cohort. The code, prompts, and reproducibility artifacts for this paper are available at https://github.com/sallyy1/CAP-ACL-2026.

Pipeline	Avg latency (s)	Rel. latency	Avg input toks	Avg output toks	Avg total toks	Rel. total toks
Direct ($X \to Y$)	10.03	1.00	2988.03	808.83	3796.87	1.00
CAP ($X \to C \to Y$)	60.71	6.06	20799.83	4255.33	25055.17	6.60
CAP+Event ($X \to C \to E \to Y$)	64.50	6.43	20634.17	4445.33	25079.50	6.61

Table 13: Deployment-path latency/token proxy on 30 cases; relative values are normalized to Direct.

D Ablation: SOAP-template Rendering

This section complements the main sectioned-template results with SOAP-template diagnostics. Table 11 reports SOAP-template aggregate metrics for the same five methods as the main experiment. Figure 4 then visualizes case-level SOAP semCAP-R win/tie/loss rates against Direct. Sectioned-template full win/tie/loss counts are reported in the main text (Table 7).

Method	semCAP-R (ours)↑	semCAP-P (ours)↑	semCAP-F1 (ours)↑	GPT-R (Nair)↑	GPT-P (Nair)↑	GPT-F1 (Nair)↑
Direct (base)	0.580	0.170	0.203	**0.636**	0.785	**0.693**
MEDSUM-ENT (prior)	0.546	0.158	0.191	0.596	0.737	0.649
Cluster2Sent (prior)	0.470	0.176	0.192	0.527	**0.791**	0.617
CAP (ours)	**0.660**	**0.210**	**0.246**	0.548	0.757	0.626
CAP+Event (ours)	0.520	0.184	0.196	0.508	0.771	0.600

Table 11: Results on the 197-case ACI-Bench cohort. (SOAP-template).

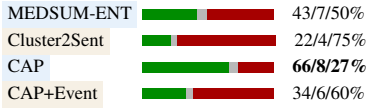

Figure 4: Case-level semCAP-R win/tie/loss rates against Direct. (SOAP-template).

E Ablation: Transcript-Free Rendering

To isolate how much of note generation can be supported by the intermediate representation alone, we evaluate two transcript-free ablations: **CAP-only** (renderer receives only the CAP set C) and **CAP+Event-only** (renderer receives C plus the bundle plan E), both without transcript X. These are not intended as competitive baselines; rather, they stress-test the *renderability* of intermediate scaffolds under a strict source constraint. Table 12 summarizes the sectioned-template means for these transcript-free conditions.

Method	semCAP-R (ours)	semCAP-P (ours)	semCAP-F1 (ours)	GPT-R (Nair)	GPT-P (Nair)	GPT-F1 (Nair)
CAP	**0.694**	0.212	0.256	**0.632**	**0.795**	**0.695**
CAP+Event	0.570	0.181	0.214	0.578	0.773	0.649
CAP-only	0.650	**0.269**	**0.277**	0.331	0.607	0.412
CAP+Event-only	0.141	0.077	0.065	0.132	0.288	0.167

Table 12: Transcript-free ablations (sectioned template; means). CAP+Event-only is a negative-control-style condition; CAP-only remains substantially stronger, indicating that proposition-level content can carry useful grounding signal even without transcript phrasing.

Interpretation. CAP-only keeps substantial grounded signal, but without transcript phrasing it underperforms on note-level quality. CAP+Event-only is much weaker, indicating that the current compact bundle plan is not yet sufficient as a standalone generation substrate.

F Inference Cost Analysis

To quantify the practical overhead of structure-first generation, we measure latency and token usage on a 30-case stratified subset using the same backbone model and sectioned-note template as the main experiments. We compare the deployment-time generation paths for Direct ($X \to Y$), CAP ($X \to C \to Y$), and CAP+Event ($X \to C \to E \to Y$). Reference-note CAP extraction is excluded because it is used only for offline evaluation, not deployment-time generation.

As shown in Table 13, the multi-stage pipelines increase both latency and token usage relative to Direct. CAP+Event adds only modest overhead over CAP, suggesting that most of the additional cost comes from CAP extraction and scaffold-conditioned rendering rather than event consolidation. These results quantify the current engineering cost of source-grounded structure-first generation and motivate future work on more efficient CAP induction, caching, and lightweight consolidation.

Case	Boundary regime	CAP snippet	CAP+Event snippet	Takeaway
D2N136	Organization gain (with residual risk)	"Continue metformin 500 mg twice daily" but "follow-up in two weeks" was flagged as unsupported by a reviewer.	Knee-pain problem and action items are grouped coherently (autoimmune labs + PT + meloxicam), but additional plan details (e.g., lipid profile, BP-cuff request) may over-extend beyond core focus.	Bundling can improve problem-level structure while still requiring tighter action filtering.
D2N203	Mixed trade-off	Captures vision-change workup and monthly Lucent injections, but carries restaurant narrative detail into HPI.	Maintains coherent AMD plan and follow-up cadence, yet can still retain low-value contextual detail from dialogue.	Consolidation helps organization, but noise suppression is still case dependent.
D2N010	Boundary failure	Hallucinates infection-heavy interpretation (Lyme/strep emphasis) and misses key CHF-related nuance in reviewer comments.	Over-compressed plan with omission risk (reviewers flagged missing hyperglycemia severity/right-leg edema and orthopnea-related nuance).	When evidence is noisy and multi-threaded, current event consolidation can under-specify critical findings.

Table 16: Clinician-reviewed boundary examples with generated-note snippets (sectioned template). Green marks clinically useful preservation; red marks reviewer-flagged or boundary-prone behavior.

G Boundary Analysis for Event Consolidation

Stratifying by bundle count. We stratify CAP+Event outputs by `event_plan_count` as a diagnostic proxy for consolidation strength (not a gold problem count). Table 15 shows that two-bundle plans are relatively safer, while one-bundle collapse is most associated with recall loss and under-generation.

`event_plan_count`	n	ΔsemCAP-R	ΔPDSQI	ΔPDSQI(org)	ΔF-CL(A+B+C)
1	144	−0.133	−0.093	−0.111	−0.191
2	37	−0.096	+0.039	−0.054	+0.027
3+	16	−0.110	−0.080	−0.062	−0.203

Table 15: Boundary analysis for event consolidation on the 197-case cohort (sectioned template). Values are mean deltas for CAP+Event relative to CAP. Bundle count is taken from the consolidation plan produced by cap_event.

Representative cases. Table 16 lists representative cases illustrating three regimes: (i) clear gains where consolidation improves organization and faithfulness (e.g., D2N207), (ii) trade-offs where organization improves but grounding recall drops (e.g., D2N190), and (iii) failure modes where consolidation under-generates or over-compresses, yielding large drops in semCAP-R and downstream quality (e.g., D2N153/D2N158). To make this concrete for readers, Table 16 additionally shows snippet-level highlights from the three clinician-reviewed cases (D2N136/D2N203/D2N010), marking preserved content in green and problematic rendering in red.

Case	Bundles	Regime	ΔsemCAP-R	ΔPDSQI	ΔPDSQI(org)	ΔF-CL(A+B+C)
D2N207	2	gain	+0.218	+2.286	+3.000	+1.000
D2N136	1	gain (org)	+0.176	+0.714	+1.000	−0.125
D2N190	2	trade-off	−0.178	+1.286	+1.000	+1.625
D2N205	1	failure	−0.886	−1.571	−2.000	−2.125
D2N153	1	failure	−0.714	−2.595	−4.000	−2.500
D2N158	1	failure	−0.675	−2.572	−3.000	−3.125

Table 16: Representative boundary cases for CAP+Event vs. CAP on the sectioned template; deltas are CAP+Event minus CAP.

H Extended Qualitative Example: D2N008

D2N008 is a multi-condition encounter in which the patient presents with fatigue and dizziness in the setting of low hemoglobin, background CHF, edema, salt-intake-related fluid retention, and a downstream GI workup plan. The main qualitative challenge is not merely whether these facts are mentioned, but whether they are organized around the correct primary problem. In our reading of the transcript, the clinically dominant structure is: *primary problem = anemia; evidence = fatigue, dizziness, low hemoglobin; context = CHF and edema; plan = anemia profile plus gastroenterology referral for endoscopy/colonoscopy*. This makes D2N008 a useful example of why proposition extraction and event construction are both needed.

The intermediate CAP representation also reveals a limitation that is useful to discuss explicitly. It captures the major signals well—fatigue, dizziness, swelling, low hemoglobin, prior CHF, and GI follow-up planning—but it also includes duplicated symptom propositions and weakly anchored low-value findings such as nasal congestion and seasonal allergies. This supports an important design conclusion: proposition-level extraction is necessary for faithful generation, but proposition normalization alone is not sufficient for producing a clinically prioritized note. A higher-order bundling step is particularly helpful when the encounter interleaves a primary workup problem with background chronic disease, historical detail, and peripheral conversational content.

Table 17: Dialogue snippets and core clinical challenges for D2N008.

Case	Transcript Key turns (abridged)	Clinical challenge
D2N008	U5: [patient] over the past couple of months , i've been really **tired** and **dizzy** . . . . U16: [doctor] okay , all right . so , you know , let's talk a little bit about that colonoscopy . i know you had a colonoscopy about three years ago and that showed that you had some mild **diverticuli- diverticulosis** . um , no issues since then ? U13: [patient] no , no weight loss or passing out . . . U19: [patient] um , it's been more of a struggle recently . i've been traveling a lot . i went up to vermont , um , to go , um , explore the mountains . and along the way i stopped at , you know , . . . i've noticed some **swelling** in my , my legs . . . U27: [patient] uh , i've had a little bit of **nasal congestion** just because with the seasons changing , i , i get seasonal allergies . but everything else has been okay . U32: [doctor] . . . physical examination looks fine . so , on your heart exam , i do hear a three out of six systolic ejection murmur , which we've heard in the past , okay ? and on your lower extremities , i do notice some trace to one plus **pitting edema** in your ankles , which is probably from the salt intake , okay ? U37: [doctor] . . . for you first problem this new anemia , uh , i wanna go ahead and send off some more labs and **anemia profile** , just to see exactly what type of anemia we're dealing with . i also wanna go and refer you back to the **gastroenterologist** for another evaluation , okay ? hey , dragon , order referral to gastroenterology . so , they're gonna do , uh , probably do an **endoscopy** and another **colonoscopy** on you . um , but again , i wanna send off those labs just to make sure that it's not something else , okay ?	• **Distributed evidence**: anemia workup depends on symptoms, prior GI history, edema, and follow-up plan scattered across the visit. • **Source grounding under noise**: low-salience seasonal-allergy content co-occurs with clinically central anemia evidence. • **Problem–plan linkage**: the anemia problem must remain linked to anemia profile, GI referral, endoscopy/colonoscopy, and medication adjustments. • **Avoiding over-attribution**: side details such as allergies should not dominate the assessment.

Table 18: Qualitative comparison on D2N008. This case illustrates distributed anemia-related evidence and a plan that must remain linked across symptoms, prior GI history, exam findings, and follow-up actions. CAP-based methods improve structural coherence, while Direct remains competitive as a strong baseline.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Low hemoglobin levels. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months, exacerbated by recent travel. She notes increased swelling in her ankles and reports difficulty keeping up with her usual activity level. She had a colonoscopy three years ago that revealed mild diverticulosis with no complications. She previously experienced kidney stones two years ago, with no recent flares. She denies any recent blood in her stools or melena. **Findings** Physical examination reveals trace to one plus pitting edema in the ankles. The patient's blood pressure is within normal limits, and her heart rate is appropriate. A three out of six systolic ejection murmur is heard on cardiac auscultation. The echocardiogram demonstrates a left ventricular ejection fraction of 45% and moderate mitral regurgitation. Hemoglobin is 8.2 g/dL. **Assessment** The patient presents with anemia, likely secondary to chronic disease or other underlying causes. Her congestive heart failure and recent diuretic use may contribute to her symptoms. **Plan** I am ordering an anemia profile to further evaluate the cause of the anemia. I am also ordering a referral to gastroenterology for an endoscopy and repeat colonoscopy. I will initiate lasix 40 milligrams once daily, continue toprol 50 milligrams daily, and lisinopril 10 milligrams daily. I will instruct the patient to monitor her weight daily and call if there is any weight gain. I will also recommend a scale to assist with monitoring sodium intake.	A strong direct baseline: it covers much of the visit correctly. Its weakness is that the overall rationale is less explicitly scaffolded, making it harder to see how distributed evidence was organized.
Cluster2Sent	**Chief Complaint** The patient reports fatigue and dizziness. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months. She denies any recent trauma. She reports some mild nasal congestion related to seasonal allergies. She denies fever, chills, or sore throat. She denies any weight loss or loss of appetite. She reports some swelling in her lower extremities, primarily in her ankles. She has no issues since her previous colonoscopy three years ago. **Findings** The patient is a 49-year-old female with a past medical history of congestive heart failure, kidney stones, and a prior colonoscopy. Her vital signs are stable, with blood pressure 132/78 and heart rate 78. A three out of six systolic ejection murmur is auscultated. The echocardiogram reveals a left ventricular ejection fraction of 45%, moderate mitral regurgitation, and trace to one plus pitting edema in the ankles. Hemoglobin is low. **Assessment** The patient presents with anemia, likely secondary to chronic disease and requiring further investigation. She also continues to have symptoms consistent with congestive heart failure. **Plan** I am going to order additional labs, including a complete blood count and iron studies. I will also refer her back to the gastroenterologist for an endoscopy and repeat colonoscopy. I will initiate lasix 40 milligrams once daily and continue toprol 50 milligrams daily. I will also consider a diuretic increase if symptoms persist.	Reasonably faithful, but flatter and less explicit about which evidence belongs to the anemia workup versus background chronic conditions.
MEDSUM-ENT	**Chief Complaint** *[empty]* **History of Present Illness** *[empty]* **Findings** *[empty]* **Assessment** *[empty]* **Plan** *[empty]*	Under-generated sectioned output (empty skeleton), limiting its usefulness as a rendering scaffold for long multi-condition narratives in our prompt-based reimplementation.

Method	Generated note (sectioned template, full)	Comment
CAP	**Chief Complaint** Low hemoglobin level. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months, particularly with increased exertion. She notes swelling in her ankles and reports difficulty with salt intake, particularly while traveling. She had a colonoscopy three years ago that revealed mild diverticulosis with no issues since. She also has a history of kidney stones two years ago with no recent flares. She reports seasonal allergies and nasal congestion. She denies any weight loss or loss of appetite, or episodes of passing blood in the stool or urine. **Findings** The patient's blood pressure is within normal limits. Heart rate is appropriate. Physical exam reveals trace to one plus pitting edema in the ankles. A systolic ejection murmur is heard. The doctor notes nasal congestion. **Assessment** The patient's low hemoglobin level is concerning and requires further investigation. Her congestive heart failure, diverticulosis, and seasonal allergies are contributing factors. **Plan** The doctor plans to order an anemia profile and a referral to the gastroenterologist for an endoscopy and colonoscopy. The doctor also orders lasix 40 milligrams once daily, toprol 50 milligrams daily, and lisinopril 10 milligrams a day. The doctor recommends monitoring weight daily and calling if it increases. The gastroenterologist will perform an endoscopy and colonoscopy.	CAP preserves the main anemia-centered plan well, but still over-carries peripheral details such as seasonal allergies into the note.
CAP + Event	**Chief Complaint** Low hemoglobin levels and fatigue. **History of Present Illness** Stephanie is a 49-year-old female with a history of congestive heart failure, kidney stones, and a prior colonoscopy. She presents today with complaints of fatigue and dizziness over the past couple of months, noting that these symptoms have worsened recently, particularly with travel. She reports increased swelling in her ankles and has been struggling to maintain a low-salt diet. She had a colonoscopy three years ago that revealed mild diverticulosis with no further issues. She also had kidney stones a couple of years ago, with no recent flares. She reports occasional nasal congestion related to seasonal allergies but no other significant symptoms, including fever, chills, sore throat, or cough. She denies any weight loss or loss of appetite and has not experienced any episodes of fainting or dark stools. **Findings** Physical exam reveals blood pressure 128/78 mmHg, heart rate 72 bpm, and trace 1+ pitting edema in the ankles. A three out of six systolic ejection murmur is heard. The echocardiogram shows a left ventricular ejection fraction of 45% and moderate mitral regurgitation. Hemoglobin is 8.2 g/dL. **Assessment** The patient presents with anemia, likely secondary to chronic disease and lifestyle factors. Her congestive heart failure and recent travel contribute to her symptoms. **Plan** I am ordering an anemia profile to determine the etiology of the anemia. I will also refer her back to the gastroenterologist for an endoscopy and repeat colonoscopy. I will initiate lasix 40 mg daily, continue toprol 50 mg daily, and lisinopril 10 mg daily. I will also order a referral to gastroenterology for further evaluation. I will instruct the patient to monitor her weight daily and call if there is any significant increase.	CAP+Event most clearly organizes the visit around the anemia problem and its downstream workup, though it can still introduce mild over-explanation in narrative form.

I Extended Qualitative Example: D2N026

D2N026 is our main safety-critical supporting case. The encounter is structurally simple compared with D2N008, but it contains a clinically important contradiction that occurs late in the dialogue: the patient first denies blood thinners and later states that he is taking Eliquis. Because the downstream plan includes cervical MRI and a tentative epidural, recovering the *final grounded medication state* is more important here than broad narrative fluency. This case therefore illustrates a second recurring failure mode in clinical dialogue summarization: brittle final-state tracking under long-range dependency. Table 19 summarizes the key dialogue turns and why this failure mode is safety-critical for downstream planning.

Table 19: Dialogue snippets and core clinical challenges for D2N026.

Case	Dialogue evidence (abridged)	Clinical challenge
D2N026	**Key turns** U9: "the thing started about **two weeks ago**" U19–U27: arm/hand pain extends down the arm; **hand weakness** and grip difficulty U49–U55: cervical source, **MRI** needed, tentative **epidural** plan U57: "do you take any **blood thinners**?" U58: "no, I don't" U59: "I am taking **Eliquis**, though" U65: "my hand is still bothering me" U66–U68: **MRI of neck + C7–T1 epidural** plan maintained	• **State update conflict**: earlier negation → later correction. • **Medication safety**: anticoagulant status matters. • **Primary problem grounding**: cervical / arm–hand pain. • **Plan preservation**: neck MRI + epidural plan.

Table 20: Qualitative comparison on D2N026. Direct generation collapses to an unrelated problem, while Cluster2Sent keeps the wrong earlier medication state. CAP-based methods better preserve the later clinically actionable Eliquis update together with MRI and epidural planning.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Right shoulder pain **History of Present Illness** The patient presents with right shoulder pain that began approximately 3 days ago after lifting a heavy box. He describes the pain as sharp and constant, exacerbated by overhead movements. He denies any recent trauma. He has tried over-the-counter ibuprofen with minimal relief. He reports difficulty reaching behind his back. He has no prior history of shoulder problems. He is currently taking lisinopril and metformin. **Findings** Physical exam reveals tenderness to palpation over the right rotator cuff. Range of motion is limited by pain. Neurological exam is intact. **Assessment** Rotator cuff strain is suspected. **Plan** Continue with ice and elevation. Initiate physical therapy. Consider a rotator cuff ultrasound to rule out a tear.	Catastrophic topic drift. The note misses the cervical MRI, epidural planning, gabapentin continuation, and anticoagulant update altogether.
Cluster2Sent	**Chief Complaint** Left arm pain **History of Present Illness** The patient reports left arm pain, initially attributed to a possible pinched nerve by a previous provider. He has been taking gabapentin and a neck pill, and reports using numbing medicine. He denies taking blood thinners. He reports some mild strain. **Findings** The patient has mild cervical restriction on examination. Cervical x-ray reveals significant disc degeneration at C5–C6 and to a lesser extent C4–C5 and C3–C4. MRI of the neck is planned to assess for nerve compression. Physical exam reveals no apparent distress. **Assessment** Suspect cervical radiculopathy with possible contributing peripheral neuropathy. The patient's presentation is consistent with cervical disc degeneration and radicular symptoms. **Plan** Order a neck MRI to assess for nerve compression. Tentatively plan for a left C7–T1 epidural, pending MRI results. Continue the patient on gabapentin and advise him to continue his home exercise program. Schedule a follow-up appointment in four weeks if symptoms persist.	Better problem anchoring than Direct, but it preserves the *wrong final state* by carrying forward the earlier blood-thinner denial instead of the later Eliquis correction.
MEDSUM-ENT	**Chief Complaint** *[empty]* **History of Present Illness** *[empty]* **Findings** *[empty]* **Assessment** *[empty]* **Plan** *[empty]*	In our prompt-based reimplementation, the sectioned renderer under-generated for this case, producing an empty skeleton. (The SOAP template produces a non-empty output but still struggles to reconcile the late Eliquis update.)
CAP	**Chief Complaint** Left arm pain. **History of Present Illness** The patient reports hand pain, which began approximately two weeks ago. He describes the pain as worsening at night and initially experienced it when waking at 2:00 AM. He reports that applying ice provided temporary relief. The patient also notes difficulty gripping things and reports weakness in his hand. He previously injured his hand and has a history of leg issues and knee issues, and is one percent disabled. He reports taking eliquis and gabapentin. He previously received numbing medicine and oral steroids. **Findings** Physical exam reveals mild cervical restriction with limited extension, and tenderness to palpation in the neck. Strength is one plus bilaterally in the biceps, triceps brachii, and brachioradialis. Reflexes are normal. Examination does not reveal any specific findings in the upper extremities other than the reported hand weakness. **Assessment** The patient's symptoms are most likely due to cervical radiculopathy, possibly secondary to disc degeneration. Peripheral neuropathy is also a consideration. **Plan** I am going to order a cervical MRI to further evaluate the nerve roots and determine the extent of compression. I will also tentatively schedule a C7–T1 epidural. The patient will continue taking gabapentin and eliquis. I will schedule a follow-up appointment in four weeks to assess his progress.	CAP recovers the clinically critical later correction (**Eliquis**) and preserves the cervical radiculopathy workup and treatment plan.

Method	Generated note (sectioned template, full)	Comment
CAP + Event	**Chief Complaint** Left arm pain. **History of Present Illness** The patient reports left arm pain that began approximately two weeks ago, worsening at night. The pain initially arose suddenly and was accompanied by a feeling of discomfort, but it has since become more persistent. He describes the pain as a dull ache, and it seems to be exacerbated by certain movements. He mentions that he experienced similar symptoms in the past, which were attributed to a pinched nerve by a previous physician. He has been taking gabapentin and eliquis as prescribed. He denies any recent trauma or injury to the arm. **Findings** Physical exam reveals mild tenderness to palpation in the cervical region. Range of motion is limited in cervical flexion and extension. Strength testing reveals a one-plus grade for biceps and triceps strength bilaterally. Reflexes are normal. Cervical x-ray demonstrates significant disc degeneration at C5–C6 and C4–C5. **Assessment** The patient's symptoms are most likely due to cervical radiculopathy, possibly exacerbated by disc degeneration. Peripheral neuropathy is also a consideration. **Plan** I am ordering a cervical MRI to further evaluate the nerve roots and determine the extent of compression. We will also schedule a left C7–T1 epidural to provide targeted anti-inflammatory relief. The patient will continue with gabapentin as prescribed and follow up in four weeks to assess symptom improvement.	CAP+Event preserves the main problem–evidence–plan chain and avoids the unsafe blood-thinner denial, though some detail is compressed during rendering.

J Extended Qualitative Example: D2N024

This case is useful precisely because it is not a clean victory. The clinically dominant structure is mild scar discomfort with conservative management and deferred revision, while the lump belongs to historical context and the review-of-systems negatives should remain suppressed. As shown in Tables 21 and 22, Cluster2Sent is brittle under negation-heavy follow-up dialogue and can promote explicitly negated symptoms into positives, while MEDSUM-ENT and Direct remain strong on this visit style. CAP contributes by making temporal resolution and the conservative scar-centered plan explicit, but it can still over-carry speculative counseling content. We therefore view D2N024 as evidence that structured representations improve robustness under noisy dialogue, but do not fully solve the filtering problem when low-salience follow-up content and speculative counseling are tightly interleaved.

Table 21: Dialogue snippets and core clinical challenges for D2N024.

Case	Dialogue evidence (abridged)	Clinical challenge
D2N024	**Key turns** U8: "**no**" to **chills**, **fever**, **nausea**, **vomiting** U18: back pain **resolved** U21: **lump under left breast** was noted in March … U23: **scars were bothering her** … U33–U47: conservative plan: **scar gel**, **massage**, photos, **revisit in 6 months**	• **Temporal resolution**: prior lump → no longer present. • **Negation-heavy follow-up**: ROS items can be falsely promoted. • **Low-salience follow-up noise**: social talk can leak into note text. • **Current problem vs. future option**: revision is future/hypothetical.

Table 22: Qualitative comparison on D2N024. This follow-up visit mixes a scar complaint with a resolved historical lump and negation-heavy review-of-systems content.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Breast scar discomfort following bilateral reduction mammoplasty. **History of Present Illness** The patient reports persistent discomfort with the scars from her bilateral reduction mammoplasty performed in October 2020. She initially reported irritation and sought relief with scar gel, applying it twice daily. She states the scars were bothersome and used the gel to mitigate this. She has not had any further issues with the scars since her last visit with Dr. Sanchez in March. She denies any new chills, fever, nausea, or vomiting. She reports enjoying summertime activities and her family is doing well. She has not experienced any new back pain. **Findings** Physical examination reveals well-healed incisions with no signs of infection or redness in either breast. The incisions appear widened, particularly on the left side. The patient reports subjective discomfort with the left scar. **Assessment** Postoperative scarring following bilateral reduction mammoplasty. **Plan** Schedule a follow-up appointment in six months to reassess the scars. Continue the use of mederma scar gel twice daily on the affected areas, massaging the scars. Obtain photographs of the scars for documentation and future comparison.	Strong baseline: keeps the scar follow-up as primary, preserves ROS negations, and treats the prior lump as historical/resolved.
MEDSUM-ENT	**Chief Complaint** Scar discomfort following bilateral reduction mammoplasty. **History of Present Illness** The patient reports ongoing discomfort with the scars from her previous bilateral reduction mammoplasty, specifically noting that one scar area is bothering her. She previously reported a lump in that area to her primary care physician, Ruth Sanchez, in March, but no longer feels it. She has been using scar gel twice daily to manage the scar tissue and has observed some widening of the incisions. She denies any signs of infection or redness. The patient is currently enjoying summertime and has not experienced any new symptoms or concerns. **Findings** Examination reveals well-healed incisions with no signs of infection or redness on either breast. The patient reports the scar tissue is widening. **Assessment** The patient's primary concern is persistent scar discomfort following breast surgery. There is no evidence of acute infection or complications. **Plan** The patient should continue using the mederma scar gel twice daily on the affected scar area. We will schedule a follow-up appointment in six months to reassess the scar tissue and determine if further intervention is needed. Photographs will be taken today to document the current state of the scars for future reference.	Captures temporal resolution (lump no longer present) and negation-heavy content cleanly; aligns with MEDSUM-ENT's status-centric design.
Cluster2Sent	**Chief Complaint** Patient reports concern regarding breast scars following bilateral reduction mammoplasty. **History of Present Illness** The patient reports experiencing chills, fever, nausea, and vomiting. She describes a lump located under her left breast, which she has not felt recently. She notes that the scars are bothering her and have widened slightly, particularly in a high-tension area. She has been using scar gel and massage to manage the scars. **Findings** The incisions are well-healed with no signs of infection or redness on either breast. **Assessment** The patient's reported breast lump and scar concerns require further evaluation. **Plan** Schedule a follow-up appointment in six months to reassess the scars and address any ongoing concerns. Recommend continued scar gel application and massage.	Promotes explicitly negated ROS items to positives (unsafe) and blurs the historical lump vs. current scar follow-up distinction.
CAP	**Chief Complaint** Patient reports scar discomfort and a desire for scar revision. **History of Present Illness** The patient reports a breast lump discovered during self-examination in March, which was evaluated by Dr. Sanchez. She reports that the scar from her previous reduction mammoplasty has been bothering her, and she has been using scar gel twice daily. She denies any recent trauma or fever. She reports no nausea or vomiting. She describes the scar as being slightly bothersome, and she's been massaging the scar gel. She's been using the gel for about a month. She's been using the gel twice a day. She discontinued using the gel because it was bothering her. She's been enjoying summertime and hasn't had any new problems. **Findings** Examination reveals well-healed incisions with no signs of infection or redness. The incisions appear widened. The patient denies chills, fever, nausea, or vomiting. **Assessment** The patient has scars from previous surgery and reports discomfort. There is no acute infection or signs of complications. **Plan** The doctor recommends the patient continue using scar gel twice a day and massaging the scar gel. The doctor recommends revisiting the scar in six months. The doctor will take pictures of the scars today and will revisit the scars in six months. The doctor recommends surgical excision if necessary.	Preserves key facts, but over-carries speculative counseling (revision/excision) and includes repetition/noise in HPI.

Method	Generated note (sectioned template, full)	Comment
CAP + Event	**Chief Complaint** Patient is Pamela Cook, a 36-year-old female post-bilateral reduction mammoplasty on 10-10-2020, presenting for follow-up regarding scar discomfort and a recent lump. **History of Present Illness** The patient reports being Pamela Cook, a 36-year-old female post-bilateral reduction mammoplasty on 10-10-2020. She states she is doing well and enjoys summertime. She reports that the scars here were kind of bothering me and I got scar gel. I was using it everyday, but I do n't think I need it now. She also mentioned a lump right here under her left breast that she felt during her mammogram in March, which was noted by Dr. Ruth Sanchez. However, she hasn't felt it since then. She discontinued using scar gel. She denies any new chills, fever, nausea, or vomiting. **Findings** Physical examination reveals well-healed incisions with no signs of infection or redness on either breast. The patient's back pain is currently not a concern. **Assessment** The patient's presentation is consistent with post-operative scar discomfort and a previously identified, now absent, palpable lesion, warranting continued monitoring. **Plan** The doctor recommends revisiting the scar in six months. The patient can keep using mederma scar gel twice a day. The patient can put scar gel on other scars. The patient should massage the scar gel. The doctor will take pictures of the scars today. The doctor will revisit the scars in six months.	Better separates historical vs. current status and keeps the conservative follow-up plan; still contains some templated repetition (name/date).

K Clinician Deep Review Protocol and Summary

We conducted a focused deep review on three sectioned-note cases: D2N136 (organization gain), D2N203 (mixed trade-off), and D2N010 (boundary failure). Two clinicians independently rated Direct, Cluster2Sent, CAP, and CAP+Event using a checklist co-designed with the reviewers to directly probe the failure modes in our Motivation and Research Questions (Figure 5). The checklist contains eight 1–5 Likert items spanning (A) content preservation, (B) clinical state fidelity, and (C) problem-oriented organization, plus safety flags (omission/hallucination/major error; yes/no). We also collected a *clinical usability* rating as an auxiliary item and report it separately (not included in the A+B+C aggregate). Overall, the three cases cover (i) an organization gain, (ii) a preservation–organization trade-off, and (iii) a boundary failure where all systems show clinically meaningful weaknesses.

Reviewer summaries.

- **Clinician A:** On close reading, notes can still contain omissions/hallucinations and occasional section-miscategorization even when they look broadly plausible.
- **Clinician B:** No severe errors were observed; differences were largely preference-level, and conclusions are limited with two reviewers.

Per-case clinician sheets. Tables 24–25 summarize the raw scoring sheets from our two clinicians (A and B) with brief English translations of their key comments. Narrative notes made Problem–Evidence Linkage difficult to rate in some cases (N/A), consistent with our decision to treat this item cautiously in the main quantitative analysis.

Clinician deep-review checklist (1–5 Likert; 1=poor, 5=excellent).

A. Content Preservation

1 (C) **Core clinical information preserved:** Are clinically salient facts from the dialogue retained without major omission?

2 (N) **Noise suppression:** Are small talk, repeated questions, and peripheral context avoided in the note?

3 (Ly) **Lay-expression fidelity:** Is patient language neither copied verbatim nor over-abstracted, preserving symptom nuance (severity/uncertainty)?

B. Clinical State Fidelity

4 (F) **State update / final-state fidelity:** Are late corrections and resolved uncertainty reflected in the final state (unknown→resolved; negation→positive; current med vs. intended change)?

5 (T) **Temporality preservation:** Are past history, current status, and future plan kept distinct?

C. Problem-oriented Organization

6 (Lk) **Problem–evidence linkage:** Are symptoms/exam/tests/history linked under the correct clinical problem?

7 (SP) **State–plan separation:** Are current state/current meds separated from new orders/prescriptions/follow-up plans?

8 (SO) **Section / note organization:** Are items placed in appropriate SOAP/sectioned sections?

Safety flags (Y/N): O=clinically meaningful omission; H=clinically concerning hallucination; M=major error.

Auxiliary: U=clinical usability (1–5; reported separately; excluded from the A+B+C faithfulness aggregate).

Figure 5: Clinician deep review rubric co-designed with the clinician reviewers to directly probe the failure modes in Motivation and Research Questions. Abbreviations match Tables 24–25.

Method	A Content	B State	C Org	Mean (A+B+C)	Clinical Usability
Direct	3.78	**5.00**	**4.89**	4.54	3.33
Cluster2Sent	4.11	**5.00**	4.72	4.63	3.67
CAP	**4.19**	**5.00**	4.72	4.62	3.33
CAP+Event	4.17	**5.00**	4.83	**4.77**	**4.00**

Table 23: Clinician deep review summary (mean across three ACI-Bench cases × two clinicians). Scores are 1–5 Likert ratings aggregated by rubric category (Figure 5); N/A items are excluded. Direct is highest on overall organization (C) in this small sample; among the two clustering/bundling approaches, CAP+Event outscored Cluster2Sent on organization and the A+B+C mean.

Case	Method	C	N	Ly	F	T	Lk	SP	SO	U	O/H/M	Key note (translated)
D2N136	Direct	3	5	N/A	N/A	5	N/A	5	5	3	Y/Y/–	Omits negated ROS items (e.g., vomiting/chest pain/dyspnea) and hallucinates a 4-week follow-up; state-update and problem–evidence linkage were not applicable to rate in this case.
D2N136	Cluster2Sent	3	5	N/A	N/A	5	N/A	5	4	3	Y/–/–	Best plan capture among the compared methods for this case, but omissions were still flagged.
D2N136	CAP	4	5	N/A	N/A	5	N/A	5	5	4	–/Y/–	Hallucinates a 2-week follow-up appointment.
D2N136	CAP+Event	3	3	N/A	N/A	5	N/A	5	3	3	Y/Y/–	Omits continued metformin (500mg BID) from the plan and hallucinates a 4-week follow-up; misses prior rotator cuff repair and over-includes low-value narrative; BP is missectioned.
D2N203	Direct	4	2	4	N/A	5	N/A	5	5	3	–/–/–	Includes conversational filler (e.g., "I will return in a few minutes") that should not appear in a clinical note.
D2N203	Cluster2Sent	5	4	3	N/A	5	N/A	5	4	4	–/Y/–	Hallucinates an "eye needle-stick" history not supported by the dialogue.
D2N203	CAP	5	3	4	N/A	5	N/A	5	5	4	–/–/–	No major issues were noted in the clinician sheet for this case/method.
D2N203	CAP+Event	5	3	4	N/A	5	N/A	5	5	4	–/–/–	No major issues were noted in the clinician sheet for this case/method.
D2N010	Direct	3	5	5	5	5	N/A	5	5	3	–/Y/–	Hallucinates a Lyme titer and a 4-week follow-up; overfocuses on side-sleeping and misses the key orthopnea-negative point (lying flat is not problematic).
D2N010	Cluster2Sent	3	4	4	5	5	N/A	5	5	3	Y/Y/–	Omits severity of hyperglycemia (blood glucose in the 300s); hallucinates/incorrectly summarizes rapid strep treated with amoxicillin and misstates alpha-gal "certainty."
D2N010	CAP	4	5	4	5	5	N/A	5	5	3	Y/–/Y	Omits right-leg edema in a CHF context; major interpretation error around sleep/orthopnea (misses that lying flat is fine).
D2N010	CAP+Event	3	5	4	5	5	N/A	5	5	3	Y/Y/–	Omits hyperglycemia severity and right-leg edema; incorrectly implies the patient avoids lying flat (dialogue indicated no orthopnea).

Table 24: Clinician A deep review sheet (translated). Scores are 1–5 Likert ratings: C=Core information, N=Noise suppression, Ly=Lay-expression fidelity, F=Final-state fidelity, T=Temporality preservation, Lk=Problem–Evidence linkage, SP=State–Plan separation, SO=Section organization, U=Overall clinical usability. O/H/M indicates whether the reviewer flagged omission/hallucination/major error (Y/–).

Case	Method	C	N	Ly	F	T	Lk	SP	SO	U	O/H/M	Key note (translated)
D2N136	Direct	2	5	3	5	5	3	5	5	3	–/–/–	Incorrect patient sex; includes an irrelevant statement about lower-extremity strength.
D2N136	Cluster2Sent	2	5	4	5	5	3	5	5	3	–/–/–	Captures the chief complaint, patient sex, and diabetes history.
D2N136	CAP	2	5	5	5	5	5	3	2	1	–/–/–	Includes content in Assessment that does not belong there.
D2N136	CAP+Event	5	5	5	5	5	5	5	5	5	–/–/–	Best overall among the four methods for this case.
D2N203	Direct	2	2	5	5	5	5	5	5	3	–/–/–	Italian-restaurant detail is unnecessary; HPI is overly long and could be condensed.
D2N203	Cluster2Sent	4	5	5	5	5	5	5	5	4	–/–/–	No additional comment provided in the sheet.
D2N203	CAP	4	2	5	5	5	5	5	5	3	–/–/–	Carries over the irrelevant "Italian restaurant" episode.
D2N203	CAP+Event	3	4	5	5	5	5	5	5	4	–/–/–	Overall clean, but seems to omit many findings.
D2N010	Direct	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	Cluster2Sent	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	CAP	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	CAP+Event	5	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.

Table 25: Clinician B deep review sheet (translated). Column definitions match Table 24. (This sheet did not include omission/hallucination/major-error flags for most rows, so O/H/M is shown as "–" throughout.)