NU_DeepHealthNLP at #SMM4H-HeaRD 2026: Entity-Conditioned Generation and a Four-Stage Pipeline for Automated SOAP Note Generation

Thanya Mysore Santhosh **Deahan Yu**
Khoury College of Computer Sciences
Northeastern University
{mysoresanthosh.th, d.yu}@northeastern.edu

Abstract

We describe two system submissions to Task 4 of the SMM4H-HeaRD 2026 Shared Task on automated SOAP note generation from doctor–patient dialogues. Our first submission is a standalone entity-conditioned generation model: Mistral-7B-Instruct-v0.1 fine-tuned with QLoRA on 8,529 MedSynth training dialogues, where both training and inference prompts include clinical entities extracted and grouped by SOAP section. Our second submission is a four-stage modular pipeline that additionally incorporates a hybrid retrieval stage and a rule-based verification stage. The key finding of this work is that incorporating structured clinical domain knowledge, in the form of NER entities grouped by SOAP section, directly into the generation prompt produces consistent and reliable improvements over dialogue-only generation. Our four-stage pipeline submission achieved an average score of 0.54 on the official test set, ranking first on the shared task leaderboard.

1 Introduction

Physicians spend between 52 and 102 minutes per day writing clinical notes (Sinsky et al., 2016), a burden that contributes to burnout and reduces time available for patient care. The SOAP note format—Subjective, Objective, Assessment, Plan—is the dominant structured documentation standard in primary care (Podder et al., 2023). Automating its generation from doctor–patient dialogues is the objective of SMM4H-HeaRD 2026 Task 4 (Organizers, 2026).

Prior work on clinical note generation includes SOAP note summarisation (Krishna et al., 2021), clinical note benchmarks (Yim et al., 2023; Abacha et al., 2023), and retrieval-augmented generation (Lewis et al., 2020). Domain-specific fine-tuning with retrieval (RAFT) (Zhang et al., 2024), combined with parameter-efficient adaptation via LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023), enables training clinical generation models at practical compute cost.

Progress on this task has been constrained by data scarcity. MedSynth (Mianroodi et al., 2025), released in 2025, provides 10,035 fully synthetic, privacy-compliant dialogue-note pairs spanning 2,001 ICD-10 codes, making it the largest open structured dialogue-to-SOAP-note dataset available.

We developed two system configurations for this task. The central design principle underlying both is *training–inference prompt alignment*: any information included in the inference prompt must also be present in the training prompt, or the fine-tuned model will degrade. The key contribution is entity-conditioned generation, in which structured clinical entities extracted from the dialogue are included in both training and inference prompts, producing the largest performance gain in our study. Source code is available at https://github.com/thanya8/dial2note.

2 Task and Data

Task 4 of SMM4H-HeaRD 2026 requires generating a structured SOAP note from a transcribed doctor–patient dialogue. The task uses the MedSynth dataset (Mianroodi et al., 2025), split by the shared task organisers into 8,529 training, 1,506 validation, and 368 test dialogues. The 368-example test set is a separate synthetic set generated using the same MedSynth pipeline but provided exclusively by the SMM4H 2026 Task 4 organisers for blind evaluation and not included in the public MedSynth release. Across all splits the total is 10,403 examples.

3 System Description

Both submissions share the same fine-tuned generation model (Stage 3) but differ in which pipeline stages are active.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 103–107

The **standalone entity-conditioned** submission uses two stages: Stage 1 (Clinical Entity Extractor), which extracts and groups clinical entities from the input dialogue by SOAP section, and Stage 3 (SOAP Writer), which generates the note conditioned on those entities and the dialogue.

The **four-stage pipeline** submission uses all four stages: Stage 1 (Clinical Entity Extractor), Stage 2 (Hybrid Retriever), which retrieves the most similar training case, Stage 3 (SOAP Writer), which generates the note conditioned on entities, the retrieved note, and the dialogue, and Stage 4 (Rule-Based Verifier), which validates and corrects the output. We describe each stage in order below.

3.1 Stage 1: Clinical Entity Extractor

Stage 1 extracts clinical entities from the input dialogue using Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) in zero-shot mode via structured prompting. We chose v0.3 for this stage because its improved instruction-following produces more reliable JSON output than v0.1 for constrained structured generation tasks.

The extraction prompt defines 10 entity types: chief complaint, symptom, diagnosis, medication, drug detail, test or lab order, vital sign, physical exam finding, referral, and follow-up instruction. The model is required to output a raw JSON array with no explanation or markdown. Post-processing validates labels, removes spans longer than seven words, and deduplicates. Extracted entities are then grouped into four SOAP-aligned categories: Subjective (chief complaint, symptom), Objective (vital sign, physical exam finding, test or lab order), Assessment (diagnosis), and Plan (medication, drug detail, referral, follow-up instruction).

Zero-shot prompt template The prompt instructs the model to return a raw JSON array with no markdown, defines 10 entity types with exact label strings (chief complaint, symptom, diagnosis, medication, drug detail, test or lab order, vital sign, physical exam finding, referral, follow-up instruction), and requires each entity to include `text`, `label`, and confidence `score` fields. The full prompt is available in the source code repository. Post-processing validates labels, normalises close variants (e.g. follow-up → follow-up instruction), removes spans longer than seven words, and deduplicates by exact match and substring containment. Inference used vLLM, temperature 0.1, top-p 0.95.

On the 1,506 validation dialogues, Stage 1 produces an average of 18.83 entities per dialogue.

3.2 Stage 2: Hybrid Retriever

Stage 2 is active only in the four-stage pipeline submission. It retrieves the single most similar training dialogue-note pair using a hybrid of BM25 sparse retrieval (Robertson et al., 1995) and FAISS-indexed dense retrieval (Johnson et al., 2021) with all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych, 2019). Ranked lists from both retrievers are fused using Reciprocal Rank Fusion (Cormack et al., 2009).

The retrieved SOAP note is included in the Stage 3 training prompt, ensuring that training and inference formats are identical (Combined NER+RAFT configuration). If the top retrieval result falls below a relevance threshold, the retriever returns nothing and Stage 3 generates from entities and dialogue alone.

3.3 Stage 3: SOAP Writer

Stage 3 is the core generation component shared by both submissions. It is Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) fine-tuned with QLoRA (Dettmers et al., 2023) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) rank 16, α 16, targeting all seven projection modules (`q`, `k`, `v`, `o`, `gate`, `up`, `down`). Rank $r = 16$ was selected following Hu et al. (2022) and Dettmers et al. (2023), introducing ≈13.6M trainable parameters (0.19% of 7B) — sufficient for this dataset size; higher-epoch runs overfitted (Table 1). Training used learning rate 2×10^{-4} with linear decay, AdamW 8-bit optimiser, effective batch size 8, and 1 epoch. We used Unsloth (Han et al., 2023) for training and vLLM (Kwon et al., 2023) for batch inference.

Entity conditioning as domain knowledge Incorporating structured clinical entities into the generation prompt proved to be the most effective design decision in this work. By grouping extracted entities according to SOAP section—Subjective, Objective, Assessment, and Plan—the prompt provides the model with explicit, section-aware clinical knowledge derived directly from the dialogue. This allows the model to ground its generation in discrete clinical facts rather than relying solely on implicit reasoning over the raw conversational text, producing the largest and most consistent performance improvement across all configurations evaluated.

Training prompt format For the standalone entity-conditioned submission, the training prompt contains extracted entities grouped by SOAP section and the input dialogue. For the four-stage pipeline submission (Combined NER+RAFT), the prompt additionally contains the retrieved SOAP note, following Retrieval-Augmented Fine-Tuning (Zhang et al., 2024). Both configurations use the identical format at inference, ensuring no training–inference mismatch.

3.4 Stage 4: Rule-Based Verifier

Stage 4 is active only in the four-stage pipeline submission. It applies rule-based quality checks across five dimensions: faithfulness (numeric values in the note grounded in the dialogue), negation (denied symptoms not reported as positive), completeness (priority NER entities present in the note), structural consistency (all four SOAP headers present), and output appropriateness (minimum length, no artefacts). Notes failing a dimension receive targeted corrections. Stage 4 requires no additional model inference calls.

4 Results

Performance is evaluated using five automatic metrics computed by the official SMM4H 2026 Task 4 evaluation script (Organizers, 2026): BLEU (Papineni et al., 2002), which measures n-gram precision between the generated and reference note; ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), which measure unigram, bigram, and longest-common-subsequence recall respectively; and METEOR (Banerjee and Lavie, 2005), which incorporates stemming and synonym matching. The official test set additionally reports ROUGE-Lsum. The leaderboard ranks submissions by average score across all five primary metrics.

4.1 Validation Set

Table 1 reports validation set performance for both configurations against published MedSynth baselines (Mianroodi et al., 2025).

The entity-conditioned model achieves 0.5464 BLEU on the validation set, exceeding the published Mistral-7B-v0.3 baseline of 0.5346. The four-stage pipeline achieves 0.5423 BLEU. Both configurations exceed all three published baselines.

4.2 Test Set and Leaderboard Ranking

Table 2 reports official test set results for both submissions. The full leaderboard is publicly available at `https://www.codabench.org/competitions/14194/#/results-tab`.

Table 1: Validation set results (1,506 samples). † = MedSynth published baselines. MTR = METEOR.

System	BLEU	R-1	R-2	R-L	MTR
GPT-4o zero-shot†	0.2616	0.6525	0.3953	0.4818	0.4754
LLaMA-3-8B†	0.5206	0.7341	0.5020	0.5740	0.6487
Mistral-7B-v0.3†	0.5346	0.7441	0.5150	0.5885	0.6589
NER-conditioned	**0.5464**	0.7454	0.5219	0.5951	**0.6645**
4-stage pipeline	0.5423	0.7454	**0.5246**	**0.5983**	0.6633

Table 2: Official test set results (368 samples) and leaderboard aggregate statistics. R-Ls = ROUGE-Lsum. MTR = METEOR.

Submission	Avg	BLEU	R-1	R-2	R-L	R-Ls	MTR
NER-conditioned	0.54	0.4415	**0.6897**	**0.4328**	**0.5139**	**0.6651**	**0.6110**
4-stage pipeline	0.54	**0.4427**	0.6866	0.4298	0.5118	0.6614	0.6103
Leaderboard aggregate (all submissions)							
Mean	0.47	0.37	0.62	0.36	0.44	—	0.55
Median	0.49	0.39	0.65	0.39	0.47	—	0.57

Table 3 shows the leaderboard ranking of our submissions relative to all other participants.

Both submissions achieved an average score of 0.54, ranking first (Table 3). The validation-to-test gap of ≈0.10 BLEU reflects distribution shift between the synthetic training data and the independently generated test set.

5 Error Analysis

We compared generated notes from the entity-conditioned model against source dialogues from the test set. The model performs well when clinical entities are explicitly stated: chief complaints, symptom durations, medication names, and explicit negations (e.g. "denies fever") are generally preserved. Multi-turn information is aggregated coherently into the History of Present Illness.

Three systematic failure modes were identified. First, the model sometimes misses secondary symptoms mentioned later in the dialogue while capturing the primary complaint. Second, temporal progression is occasionally lost: "mild for three months, sharply worsened last week" is sometimes compressed to "symptoms present for three months." Third, the model occasionally introduces assessments not directly grounded in the dialogue, suggesting that the generation model relies on learned co-occurrence patterns in addition to the explicit dialogue content.

Table 3: SMM4H-HeaRD 2026 Task 4 leaderboard. Our submission (**dy_nu**) is shown in bold. Avg = mean of BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR.

Rank	Team	Avg
1	**dy_nu (ours)**	**0.54**
2	hsiaoeric	0.53
3	bogdansmm4h	0.51
4	NoviceTrio	0.49
5	aatishp	0.48
6	kikihh	0.44
7	hlshao	0.28

6 Conclusion

We presented two system configurations for SMM4H-HeaRD 2026 Task 4. Both share an entity-conditioned QLoRA fine-tuned generation model in which clinical entities extracted from the dialogue are grouped by SOAP section and included in both training and inference prompts, eliminating the training–inference format mismatch. The four-stage pipeline additionally incorporates hybrid retrieval and rule-based verification. Our four-stage submission ranked first on the shared task leaderboard with an average score of 0.54. The central finding is that entity-conditioned generation, including NER entities grouped by SOAP section in both training and inference prompts, produced the best results in this study, outperforming dialogue-only generation and all published baselines. Additionally, limitations include evaluation on fully synthetic data, reliance on lexical-overlap metrics that correlate imperfectly with clinical quality (Moramarco et al., 2022), and absence of human clinical evaluation. Therefore, future work would benefit from richer entity supervision, including fine-tuned extractors and ontology-grounded normalisation. It would also benefit from robust evaluation methods, such as claim-level faithfulness verification and human-in-the-loop evaluation.

References

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 (EACL)*, pages 2291–2302.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Tim Dettmers and 1 others. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 10088–10115.

Daniel Han, Michael Han, and Unsloth Team. 2023. Unsloth. `https://github.com/unslothai/unsloth`.

Edward J. Hu and 1 others. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547. ArXiv preprint arXiv:1702.08734 (2017).

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 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 (ACL-IJCNLP)*, pages 3552–3562. ArXiv:2005.01795 (May 2020).

Woosuk Kwon and 1 others. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

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

Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. MedSynth: Realistic, synthetic medical dialogue-note pairs. *Preprint*, arXiv:2508.01401.

Francesco Moramarco and 1 others. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5739–5754.

SMM4H Organizers. 2026. #SMM4H-HeaRD 2026 shared task 4: Dialogue-to-note generation. https://healthlanguageprocessing.org/smm4h-2026/.

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

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2023. SOAP notes. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3), NIST Special Publication 500-226*, pages 109–126.

Christine Sinsky and 1 others. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

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. *Scientific Data*, 10(1):586.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. *arXiv preprint arXiv:2403.10131*.