MedMind AI at #SMM4H-HeaRD 2026: Data Extraction and Generation Using Prompt Engineering and Structured Outputs (Tasks 1–6)

Aatish Pradhan
Independent Researcher
aatishpradhan11@gmail.com

Brian Habersberger
Independent Researcher
habersberger@gmail.com

Abstract

Six tasks from the SMM4H–HeaRD 2026 workshop were addressed with task-specific large-language-model (LLM) pipelines relying on prompt engineering, strict structured (JSON) responses, and deterministic rule sets. The pipelines utilize no task-specific fine-tuning and can be adapted across diverse clinical and social media data. This study demonstrates that general-purpose LLMs (**gpt-5.4-mini** and **gpt-5.4**) can accurately extract and classify crucial health information when constrained by strict output schemas. Notably, our hybrid approach achieved the best overall performance among all participating systems for Task 2 (Insomnia Detection).

1 Introduction

Social-media mining for health (SMM4H) tasks conventionally employ domain-specific BERT variants or sequence-to-sequence fine-tuning. These methods incur substantial computational overhead and require extensive supervised datasets.

Recent large language models (LLMs) mitigate this via constrained inference, compelling the model to emit syntactically valid JSON that follows a user-supplied schema. Embedding key decision logic within this schema often circumvents the need for explicit task-specific fine-tuning (Pradhan et al., 2025).

We evaluate this approach across six heterogeneous tasks for SMM4H–HeaRD 2026 (Lopez-Garcia et al., 2026): multilingual adverse drug events, insomnia extraction in MIMIC-III (Johnson et al., 2016), flu vaccination status (Xu et al., 2024), Dial2Note Subjective, Objective, Assessment, and Plan (SOAP) generation (Mianroodi et al., 2025), patient metadata detection (Klein et al., 2025), and Tumor, Node, and Metastasis (TNM) pathology staging (Kefeli et al., 2024). The main contributions are: 1. A schema-constrained framework (**gpt-5.4-mini**, **gpt-5.4**) leveraging targeted prompt engineering to address six distinct tasks without fine-tuning. 2. Demonstrating that hybrid LLM-regex pipelines improve clinical extraction fidelity (Task 2). 3. Analyzing discrepancies between ground-truth labels and LLM predictions through the lens of strict entity-relationship schemas.

2 Methodology

Structured output schemas guide the model to generate responses that are subsequently parsed or processed using rule-based label assignment. By formalizing both the entities and their relationships to source evidence through this schema, the LLM can reliably disentangle complex narratives and deliver machine-readable extractions aligned with task guidelines.

2.1 Architecture

- **Input:** Raw text from the provided social media or clinical datasets.
- **Prompt Assembly:**
  - (SYSTEM) high-level domain instruction;
  - (USER) the cleaned document;
  - (SCHEMA) strict JSON output definition.
- **Model Call:** The majority of tasks utilize **gpt-5.4-mini** at temperature 0.0 with strict JSON structured outputs enforced. For the multilingual dataset (Task 1), the larger **gpt-5.4** model was deployed to handle the increased complexity of cross-lingual reasoning. Average inference time per document was approximately 1.2 seconds. Across the entire project lifecycle (totaling 132,177 requests across training, validation, and evaluation), the empirical API cost averaged $0.0024 per sample.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 187–205

- **Validation:** Rule-based heuristics were applied to the structured LLM responses to map findings to final task labels.

Prompt architectures and schema constraints were iteratively refined during the development phase, with final configurations selected based on their empirical performance on the validation sets.

2.2 Task 1: Multilingual ADE Detection

Task 1 centers on extracting adverse drug effects across multiple languages. Review of the training data revealed language- and source-specific annotation guidelines. For example, in the German dataset, dietary supplements and herbal/phytomedicines were annotated as pharmaceutical drugs, requiring an expanded definition within the prompt. Furthermore, brand name disambiguation was critical; in the Japanese dataset, "ルル" (Lulu) could refer to either a common cold medicine or a pet name. Language-specific system prompts were engineered to address these observed exceptions, improving the resolution of adverse drug events without post-hoc rule processing.

2.3 Task 2: Insomnia in Clinical Notes

Task 2 required identifying diagnostic criteria and medication spans in clinical notes. The architecture incorporated a hybrid approach: a deterministic regex-based annotator first flagged criteria and canonical drug names. This draft was passed to the LLM for contextual review to make potential corrections.

2.4 Task 3: Flu Vaccine Status on Twitter

This classification task differentiates flu vaccination intentions, actions, and test results. The prompt included the tweet's `created_at` timestamp to provide temporal grounding, ensuring the LLM could distinguish between current and previous flu seasons.

2.5 Task 4: Dial2Note Generation

Task 4 focuses on generating structured SOAP notes from clinical dialogue. Strict grounding rules were provided in the prompt, steering the model from fabricating demographics or provider names not explicitly present in the source text.

2.6 Task 5: Patient Metadata Detection

The objective is to perform a binary classification distinguishing online article sentences that report patient-related metadata from those that do not. The schema was simplified to a binary metadata flag, successfully preventing the model from falling into negative classification traps.

2.7 Task 6: TNM Extraction

This task extracts T, N, and M cancer stages from pathology reports. Engineered search directives within the prompt forced the model to actively scan for distant organ involvement, preventing it from reflexively defaulting to an M0 stage.

3 Structured LLM Outputs as a Lens for Task Definition

High-performance manual prompt engineering, coupled with a strictly enforced structured response format, improves accuracy while revealing subtle variations in the interpretation of clinical texts.

Task 1 Validation Insights: Handling multilingual extraction highlighted how schemas expose cultural task ambiguities. Applying a uniform set of English-centric extraction rules misclassified phytomedicines in the German dataset (where they are often annotated as pharmaceutical drugs). Moving these constraints into language-specific system prompts allowed the LLM to apply rules with contextual nuance.

Task 2 Validation Insights: Extracting evidence spans for insomnia highlighted the challenges of rigid token matching. The LLM naturally extracts broader semantic contexts (e.g., a full explanatory sentence) rather than the minimal specific phrases preferred by annotators. Utilizing a hybrid deterministic-LLM approach proved essential for balancing semantic understanding with strict extraction boundaries.

Task 3 Validation Insights: Initial LLM outputs frequently struggled to distinguish between past actions, current actions, and future intentions (e.g., misclassifying future intentions as "Currently-Vaccinated"). Dynamically injecting the `created_at` timestamp anchored the LLM's reasoning in a concrete temporal context.

Task 4 Validation Insights: When generating SOAP notes, the LLM initially hallucinated clinical content. This discrepancy was traced back to an over-reliance on few-shot examples in the prompt, which inadvertently anchored the model to specific formatting patterns from the training data (which contained baseline dataset hallucinations). Removing few-shot examples and enforcing

explicit instruction-based grounding yielded vastly more faithful outputs.

Task 5 Validation Insights: The interaction between the JSON schema and reasoning revealed a phenomenon we term "trap commitment bias." We formally define this as the phenomenon where an LLM, once forced by a schema to explicitly categorize a challenging negative example (the "trap," e.g., virology methodology), anchors its subsequent token generation to that negative classification, causing it to ignore secondary positive signals within the same text. This differs from standard anchoring bias and contextualizes the model's resistance to correction noted in the baseline system (Klein et al., 2025). Simplifying to a boolean output resolved this discrepancy.

Task 6 Validation Insights: Extracting TNM cancer stages revealed a strong LLM bias toward majority classes, particularly assuming no metastasis ($M0$). Engineering strict search directives within the prompt forced the model to explicitly document its search for distant metastases in a chain-of-thought field prior to emitting the final stage.

4 Results

Table 1 summarizes the performance of our system across all six tasks on the official SMM4H-HeaRD 2026 test sets.

Our strictly schema-constrained approach achieved highly competitive results, most notably yielding the best overall performance among all participating systems for Task 2 (Insomnia Detection). In Subtask 1, our hybrid LLM-regex pipeline achieved an F1-score of 0.8649. In Subtask 2, our system secured an Exact Match span extraction score of 0.7170 and a Label Classification F1 of 0.8764, substantially outperforming the respective task means. In Task 1, our unweighted macro F1-score of 0.6518 exceeded the participant mean. Temporal grounding in Task 3 enabled us to surpass participant means across both subtasks. For Task 6, our system nearly achieved perfect extraction on the primary Test Set 1. The system performed slightly above average for Task 4, and underperformed on Task 5.

5 Discussion

The performance metrics across the test sets validate the core hypothesis that schema-constrained, general-purpose LLMs can effectively compete with, and occasionally outperform, task-specific fine-tuned models in clinical NLP.

The Superiority of Hybrid Systems for Span Extraction: The system's best-in-class performance on Task 2 underscores a known limitation of pure LLM approaches: while they excel at semantic reasoning, they frequently fail at precise, token-level boundary extraction. The success of our pipeline resulted from delegating rigid span boundary identification to a deterministic regex annotator and utilizing the LLM solely as a contextual filter.

Limitations of Zero-Shot Schemas in Specialized Domains: Conversely, the underperformance on Task 5 (Metadata Detection) highlights a challenge of this approach. Although simplifying the schema to mitigate "trap commitment bias" improved our validation metrics, reasoning against our prompt was insufficient to represent the complex relationship between the epidemiological data and annotations.

The Necessity of Grounded Reasoning: Tasks 3 and 6 demonstrated that LLMs are highly susceptible to their pre-training priors without explicit grounding. Injecting temporal anchors (Task 3) and forcing the model to explicitly document its search for metastasis in a chain-of-thought field (Task 6) successfully prevented the model from defaulting to majority classes.

Evaluation Metrics vs. Clinical Factuality: In Task 4, our prompt enforced strict clinical grounding, explicitly prohibiting the model from fabricating demographic details not present in the source dialogues. However, the reference "ground truth" notes frequently contain these hallucinated demographic details resulting from the dataset's dialogue-to-notes synthesis process. Standard automated metrics like BLEU penalize strict clinical factuality when rewarding baseline hallucinations.

Limitations

A primary limitation of this work is the highly task-specific nature of the prompts and schemas. While the underlying strategy of schema-constrained inference is highly adaptable and requires very low computational overhead to initiate, the specific pipelines and logical rule sets described here do not generalize zero-shot to new tasks without manual dialogue with the model, adjustment of prompts, and schema redesign. Additionally, the approach relies on proprietary, closed-weight models (GPT-

Task	Evaluation Target	Primary Metric	Val (Ours)	Test (Ours)	Test Mean	Organizer Baseline
1	Multilingual ADE	Macro F1	0.5923	**0.6518**	0.5465	-
2, ST1	Binary Classification	F1-score	0.9091	0.8649	0.6805	**0.9300**
2, ST2	Label Classification	Micro F1	0.8000	0.8764	0.5888	**0.9300**
2, ST2	Span Extraction	Exact Match F1	0.5974	**0.7170**	0.3129	-
3, ST1	Flu Vaccination	Micro F1	0.8148	**0.8932**	0.8499	0.8700
3, ST2	Flu Test Result	Micro F1	0.9704	**0.9326**	0.9071	0.8700
4	Text Generation	Average Score	0.5203	0.4800	0.4700	**0.6082**
5	Metadata Detection	F1-score	0.5083	0.5860	0.7290	**0.7760**
6 (TS1)	TNM Staging (T)	Macro F1	0.8438	**0.9960**	-	0.9920
6 (TS1)	TNM Staging (N)	Macro F1	0.8483	**0.9980**	-	0.7830
6 (TS1)	TNM Staging (M)	Macro F1	0.7543	**0.9970**	-	0.7960
6 (TS2)	TNM Staging (T)	Macro F1	-	**0.8490**	-	0.4540
6 (TS2)	TNM Staging (N)	Macro F1	-	**0.8650**	-	0.5910
6 (TS2)	TNM Staging (M)	Macro F1	-	**1.0000**	-	0.5540

Table 1: Consolidated Performance Matrix. Validation scores reflect internal evaluations prior to submission. "Test Mean" represents the average performance of participating systems, while "Organizer Baseline" represents the organizer-provided reference models. Bold values indicate the highest score across our system, the participant mean, and the baseline.

5.4 series), limiting full transparency into the model's underlying pre-training priors.

6 Conclusion

This work introduced a framework using structured output schemas combined with optimized manual prompt strategies to extract vital health information from unstructured texts. By embedding entity-relationship frameworks into schema-constrained prompts, researchers can guide LLMs to maintain a stable chain of reasoning and achieve highly competitive task performance without fine-tuning.

References

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.

J. Kefeli, J. Berkowitz, J. M. Acitores Cortina, and 1 others. 2024. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nature Communications*, 15:8916.

Ari Z. Klein, Davy Weissenbacher, Karen O'Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models. *medRxiv*.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

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.

Aatish Pradhan, Brian M. Habersberger, James H. Wade, Denver Dsouza, and Nihal Paul. 2025. LLM pros at SMM4H-HeaRD 2025 data extraction using prompt engineering and structured outputs (task 1, 2, 3, 4, 5, 6). In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. Workshop: #SMM4H-HeaRD 2025: Joint 10th Social Media Mining for Health and Health Real-World Data Workshop and Shared Tasks.

D. Xu, G. L. García, K. O'Connor, H. Holston, A. Z. Klein, I. F. Amaro, M. Scotch, and G. Gonzalez-Hernandez. 2024. Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting. In *AMIA Annual Symposium Proceedings*, pages 1404–1413. American Medical Informatics Association.

A Appendix: Detailed Pipeline Elements

A.1 Task 1: Language-Specific Validation Performance

To contextualize the macro F1-score, the model's performance on the validation split for Task 1 is broken down by language below:

- **English (F1-en):** 0.6577
- **French (F1-fr):** 0.7059
- **German (F1-de):** 0.5577
- **Russian (F1-ru):** 0.5650
- **Overall (Macro F1):** 0.5923

A.2 Task 2: Regex-Based Annotator Details

To overcome the LLM's tendency to extract overly broad semantic spans, a deterministic Regex annotator was deployed as a first pass. The annotator used a predefined dictionary of exact-match string arrays covering all Group A and Group B canonical drug names, as well as highly specific diagnostic criteria phrasing mapped directly from the MIMIC-III training data. The outputs of this deterministic pass were then ingested by the LLM, which acted purely as a pruning filter to eliminate false-positive span extractions based on the surrounding sentence context.

A.3 Task 1: Representative Language-Specific Prompts and Schema

To address source-specific annotation guidelines, the systemic domain rules were tailored per language. Below are excerpts demonstrating how the pharmaceutical definition was adapted between the English and German prompts prior to schema enforcement:

English System Prompt Excerpt:

```
// SYSTEM PROMPT
You are an expert at detecting Adverse Drug Events (
    ADEs) in English social media posts from
    Twitter/X.
DEFINITION: An ADE is an unwanted, harmful, or
    unpleasant health effect that someone ACTUALLY
    EXPERIENCED as a result of taking a
    pharmaceutical drug or medication.
Rule 1: A pharmaceutical drug or OTC medicine is
    mentioned or clearly implied (NOT supplements,
    vitamins, coffee, alcohol, energy drinks, food,
     or herbal remedies).
```

German System Prompt Excerpt:

```
// SYSTEM PROMPT
You are an expert at detecting Adverse Drug Events (
    ADEs) in German patient forum posts.
FOR GERMAN DATA - EXPANDED PHARMACEUTICAL DEFINITION
    :
All of the following count as drugs/medications:
- Prescription drugs and OTC medicines
- Dietary supplements and vitamins (vitamin D,
    magnesium, omega-3, etc.)
- Herbal medicines and phytomedicines (black cohosh/
    Traubensilberkerze, St. John's Wort/
    Johanniskraut, valerian, agnus castus)
- Hormone preparations including HRT (Estragest,
    Famenita, Zoladex, Livial, hormonal IUD)
```

JSON Schema:

```
{
  "name": "adverse_effect_detection",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "unique_id",
      "language",
      "english_version",
      "reasoning",
      "adverse_drug_event_mentioned",
      "verbatim_ADE_mention_list"
    ],
    "additionalProperties": false,
    "properties": {
      "unique_id": {
        "type": "string",
        "description": "The unique identifier as
            provided in the text"
      },
      "language": {
        "type": "string",
        "description": "The language of the text",
        "enum": [
          "English", "German", "French", "Russian",
          "Mandarin", "Japanese", "Persian", "Other"
        ]
      },
      "english_version": {
        "type": "string",
        "description": "English translation of the
            text. Repeat the original if already in
             English."
      },
      "reasoning": {
        "type": "object",
        "description": "Step-by-step reasoning
            fields to evaluate before making the
            final classification.",
        "required": [
          "drug_mentioned",
          "drug_name_if_any",
          "is_pharmaceutical",
          "symptom_described",
          "symptom_summary",
          "who_experienced",
          "is_personal_experience",
          "temporal_relationship",
          "is_therapeutic_effect",
          "classification_rationale"
        ],
        "additionalProperties": false,
        "properties": {
          "drug_mentioned": {
            "type": "boolean",
            "description": "Is any drug or
                medication mentioned or clearly
                implied?"
          },
          "drug_name_if_any": {
            "type": "string",
            "description": "Name of the drug(s)
                mentioned, or 'none' if no drug is
                mentioned."
          },
          "is_pharmaceutical": {
            "type": "boolean",
            "description": "Is the substance a
                pharmaceutical drug or OTC medicine
                ? False for coffee, alcohol, food,
                supplements (unless German data
                where supplements count)."
          },
          "symptom_described": {
            "type": "boolean",
            "description": "Is an unintended health
```

```
            effect or symptom described in the
            post?"
      },
      "symptom_summary": {
        "type": "string",
        "description": "Brief summary of the
            symptom(s), or 'none' if no
            symptoms described."
      },
      "who_experienced": {
        "type": "string",
        "description": "Who experienced the
            symptom?",
        "enum": [
          "the author", "author's child/
              dependent", "a friend/family
              member",
          "someone else", "people in general", "
              no one (hypothetical)", "unknown"
        ]
      },
      "is_personal_experience": {
        "type": "boolean",
        "description": "Did the author or
            someone they know (family, friend)
            actually experience this effect?
            True for first-hand AND witnessed
            experiences. False for third-party
            news, general warnings,
            hypotheticals."
      },
      "temporal_relationship": {
        "type": "string",
        "description": "What is the relationship
             between the drug and the symptom
            ?",
        "enum": [
          "drug caused symptom", "drug failed to
               help (symptom persisted)",
          "drug relieved symptom", "symptom
              preceded drug", "no clear link",
              "not applicable"
        ]
      },
      "is_therapeutic_effect": {
        "type": "boolean",
        "description": "Is the described effect
            the intended therapeutic action of
            the drug (e.g., pain relief from
            painkiller, sleep from sleeping
            pill)?"
      },
      "classification_rationale": {
        "type": "string",
        "description": "One sentence explaining
            why this is or is not an ADE,
            referencing the reasoning fields
            above."
      }
    }
  },
  "adverse_drug_event_mentioned": {
    "type": "boolean",
    "description": "Final classification: true
        if a pharmaceutical drug caused a
        negative effect that was actually
        experienced by the author or someone
        they describe (first-hand or witnessed)
        ."
  },
  "verbatim_ADE_mention_list": {
    "type": "array",
    "description": "If
        adverse_drug_event_mentioned is true,
        cite the specific verbatim evidence. If
         false, this must be an empty list.",
    "items": {
      "type": "object",
      "additionalProperties": false,
      "required": ["drug_name", "original", "
          english"],
      "properties": {
        "drug_name": {
          "type": "string",
          "description": "The exact name of the
              drug as mentioned in the text"
        },
        "original": {
          "type": "string",
          "description": "The verbatim ADE
              mention in the original language"
        },
        "english": {
          "type": "string",
          "description": "The verbatim ADE
              mention in English"
        }
      }
    }
  }
}
```

A.4 Task 2: Insomnia Prompts and Schema

Prompt Structure:

```
// SYSTEM PROMPT
You are an assistant specializing in identifying
    different insomnia-related diagnostic criteria
    in anonymized clinical notes. Use the
    definitions below to apply classifications to
    the notes provided by the user.

<Definition: Difficulty Sleeping>
The patient is considered to have difficulty
    sleeping if they report any of the following:
Trouble initiating sleep.
Trouble maintaining sleep.
Waking up earlier than desired.
An explicit mention of insomnia.
</Definition>

<Definition: Daytime Impairment>
The patient is considered to have daytime impairment
     if they report any of the following:
Fatigue or malaise.
Impaired attention, concentration, or memory.
Impaired social, family, occupational, or academic
    performance.
Mood disturbance or irritability.
Daytime sleepiness.
Behavioral problems such as hyperactivity,
    impulsivity, or aggression.
Decreased motivation, energy, or initiative.
Proneness to errors or accidents.
Concerns or dissatisfaction with sleep.
</Definition>

<Drugs in Drug Group A>
Estazolam: sold under the brand name Prosom, Esilgan
    , Eurodin, Nuctalon, & others, is a
    tranquilizer medication of the
    triazolobenzodiazepine (TBZD) class...
Eszopiclone: sold under the brand name Lunesta,
    Eszop, & others, is a medication used in the
    treatment of insomnia. Evidence supports slight
     to moderate benefit up to six months. It is
    taken by mouth.
Flurazepam: (marketed under the brand names Dalmane,
     Dalmadorm, Fluzepam) is a drug which is a
    benzodiazepine derivative...
Lemborexant: sold under the brand name Dayvigo, is
    an orexin antagonist medication which is used
    in the treatment of insomnia. It is indicated
    specifically for the treatment of insomnia
    characterized by difficulties with sleep onset
    and/or maintenance in adults...
Quazepam: sold under the brand name Doral among
    others, is a relatively long-acting
    benzodiazepine derivative drug developed by the
     Schering Corporation in the 1970s. Quazepam is
     used for the treatment of insomnia...
Ramelteon: sold under the brand name Rozerem among
    others, is a melatonin agonist medication which
     is used in the treatment of insomnia...
Suvorexant: sold under the brand name Belsomra, is
    an orexin antagonist medication which is used
    in the treatment of insomnia...
Temazepam: sold under the brand name Restoril,
    Normison, Nortem, & others, is a medication of
    the benzodiazepine class which is generally
```

```
    used to treat severe or debilitating insomnia
    ...
Triazolam: sold under the brand name Halcion among
    others, is a central nervous system (CNS)
    depressant tranquilizer of the
    triazolobenzodiazepine (TBZD) class...
Zaleplon: sold under the brand name Sonata among
    others, is a sedative and hypnotic which is
    used to treat insomnia. It is a
    nonbenzodiazepine or Z-drug of the
    pyrazolopyrimidine class...
Zolpidem: sold under the brand name Ambien among
    others, is a medication primarily used for the
    short-term treatment of sleeping problems...
</Drugs in Drug Group A>

<Drugs in Drug Group B>
Acamprosate: sold under the brand name Campral, is a
     medication which reduces alcoholism cravings.
Alprazolam: sold under the brand name Xanax among
    others, is a fast-acting, potent tranquilizer
    of moderate duration within the
    triazolobenzodiazepine group...
Clonazepam: sold under the brand name Klonopin,
    Rivotril, Paxam, & others, is a benzodiazepine
    medication used to prevent and treat anxiety
    disorders, seizures, bipolar mania...
Clonidine: sold under the brand name Catapres among
    others, is an a2A-adrenergic receptor agonist
    medication used to treat high blood pressure,
    attention deficit hyperactivity disorder (ADHD)
    ...
Diazepam: sold under the brand name Valium among
    others, is a medicine of the benzodiazepine
    family that acts as an anxiolytic
Diphenhydramine: sold under the brand name Benadryl,
     Unisom, Nytol, & others, is an antihistamine
    and sedative.
Doxepin: sold under the brand name Sinequan,
    Quitaxon, Aponal, & others, is a medication
    belonging to the tricyclic antidepressant (TCA)
    [10] class...
Gabapentin: sold under the brand name Neurontin
    among others, is an anticonvulsant medication
    primarily used to treat neuropathic pain...
Hydroxyzine: sold under the brand names Atarax and
    Vistaril among others, is an antihistamine
    medication
Lorazepam: sold under the brand name Ativan among
    others, is a benzodiazepine medication
Melatonin: sold under the brand name Circadin,
    Slenyto, & others, is a naturally occurring
    hormone produced in the brain that is also used
     as a dietary supplement...
Mirtazapine: sold under the brand name Remeron among
     others, is an atypical tetracyclic
    antidepressant...
Olanzapine: sold under the brand name Zyprexa among
    others, is an atypical antipsychotic primarily
    used to treat schizophrenia and bipolar
    disorder
Quetiapine: sold under the brand name Seroquel among
     others, is an atypical antipsychotic
    medication used in the treatment of
    schizophrenia, bipolar disorder, bipolar
    depression, and major depressive disorder
Trazodone: sold under the brand name Desyrel,
    Trittico, & others, is an antidepressant
    medication used to treat major depressive
    disorder, anxiety disorders, and insomnia.
</Drugs in Drug Group B>

// USER PROMPT
Analyze the following anonymized clinical note and
    identify all evidence relevant to the insomnia
    diagnostic criteria defined in the system
    prompt.

Instructions:
1. Read the note carefully.
2. For each criterion under "Difficulty Sleeping" (
    Definition 1) that is supported by the note,
    add an entry to `
    difficulty_sleeping_identified_criteria` with:
   - the matching `criterion` (from the allowed enum
       ),
   - a minimal `citation` quoted from the note (use
       "..." to indicate omitted text within a
       citation),
   - a `regular_expression_matching_citation` that
       uniquely matches that citation in the
       original note.
3. Do the same for each criterion under "Daytime
    Impairment" (Definition 2) in `
    daytime_impairment_identified_criteria`.
4. For every Group A drug (primary insomnia
    medication) mentioned as prescribed/
    administered to the patient, add an entry to `
    group_a_drugs` with the drug name (generic form
     from the enum), citation, and regex.
5. For every Group B drug (secondary medication)
    mentioned as prescribed/administered, add an
    entry to `group_b_drugs` likewise. Recognize
    brand names and map them to the generic enum
    value.
6. Only include criteria and drugs that are
    explicitly supported by the note. Do not infer
    or speculate. If nothing applies for a category
    , return an empty array.
7. Citations must be copied verbatim from the note (
    aside from "..." for elision) so they can be
    located by string/regex search.
8. Return ONLY the JSON object matching the required
     schema - no prose, no markdown.
```

JSON Schema:

```
{
  "name": "clinical_notes_criteria",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Brief step-by-step reasoning
             grounded in the note before producing
            structured output."
      },
      "difficulty_sleeping_identified_criteria": {
        "type": "array",
        "items": {
          "type": "object",
          "required": [
            "criterion",
            "citation",
            "regular_expression_matching_citation"
          ],
          "additionalProperties": false,
          "properties": {
            "criterion": {
              "type": "string",
              "enum": [
                "Trouble initiating or maintaining
                     sleep",
                "Waking up earlier than desired",
                "An explicit mention of insomnia"
              ]
            },
            "citation": {
              "type": "string",
              "description": "Cite a concise and
                   minimal amount of text (indicate
                   sub- and partial sentence
                   sampling via '...') supporting
                   the identified criterion. For
                   example, 'Reporting being unable
                   to sleep...' is a concise and
                   minimal citation supporting '
                   Trouble initiating or maintaining
                    sleep'"
            },
            "regular_expression_matching_citation":
                {
              "type": "string",
              "description": "A regular expression
                   that can be used to uniquely
                   match the citation"
            }
          }
        }
      },
      "daytime_impairment_identified_criteria": {
```

```
  "type": "array",
  "items": {
    "type": "object",
    "required": [
      "criterion",
      "citation",
      "regular_expression_matching_citation"
    ],
    "additionalProperties": false,
    "properties": {
      "criterion": {
        "type": "string",
        "enum": [
          "Fatigue or malaise",
          "Impaired attention, concentration,
               or memory",
          "Impaired social, family,
               occupational, or academic
               performance",
          "Mood disturbance or irritability",
          "Daytime sleepiness",
          "Behavioral problems such as
               hyperactivity, impulsivity, or
               aggression",
          "Decreased motivation, energy, or
               initiative",
          "Proneness to errors or accidents",
          "Concerns or dissatisfaction with
               sleep",
          "An explicit mention of insomnia"
        ]
      },
      "citation": {
        "type": "string",
        "description": "Cite a minimal amount
             of text (indicate sub- and
             partial sentence sampling via
             '...') supporting the identified
             criterion"
      },
      "regular_expression_matching_citation":
          {
        "type": "string",
        "description": "A regular expression
             that can be used to uniquely
             match the citation"
      }
    }
  }
},
"group_a_drugs": {
  "type": "array",
  "description": "List of drugs classified
       under group A.",
  "items": {
    "type": "object",
    "properties": {
      "drug_name": {
        "type": "string",
        "enum": [
          "Estazolam",
          "Eszopiclone",
          "Flurazepam",
          "Lemborexant",
          "Quazepam",
          "Ramelteon",
          "Suvorexant",
          "Temazepam",
          "Triazolam",
          "Zaleplon",
          "Zolpidem"
        ]
      },
      "citation": {
        "type": "string",
        "description": "Cite a minimal amount
             of text (indicate sub- and
             partial sentence sampling via
             '...') supporting the identified
             drug"
      },
      "regular_expression_matching_citation":
          {
        "type": "string",
        "description": "A regular expression
             that can be used to uniquely
             match the citation"
      }
    },
    "required": [
      "drug_name",
      "citation",
      "regular_expression_matching_citation"
    ],
    "additionalProperties": false
  }
},
"group_b_drugs": {
  "type": "array",
  "description": "List of drugs classified
       under group B.",
  "items": {
    "type": "object",
    "properties": {
      "drug_name": {
        "type": "string",
        "enum": [
          "Acamprosate",
          "Alprazolam",
          "Clonazepam",
          "Clonidine",
          "Diazepam",
          "Diphenhydramine",
          "Doxepin",
          "Gabapentin",
          "Hydroxyzine",
          "Lorazepam",
          "Melatonin",
          "Mirtazapine",
          "Olanzapine",
          "Quetiapine",
          "Trazodone"
        ]
      },
      "citation": {
        "type": "string",
        "description": "Cite a minimal amount
             of text (indicate sub- and
             partial sentence sampling via
             '...') supporting the identified
             drug"
      },
      "regular_expression_matching_citation":
          {
        "type": "string",
        "description": "A regular expression
             that can be used to uniquely
             match the citation"
      }
    },
    "required": [
      "drug_name",
      "citation",
      "regular_expression_matching_citation"
    ],
    "additionalProperties": false
  }
},
"required": [
  "reasoning",
  "difficulty_sleeping_identified_criteria",
  "daytime_impairment_identified_criteria",
  "group_a_drugs",
  "group_b_drugs"
],
"additionalProperties": false
}
}
```

A.5 Task 3: Flu Vaccine Prompts and Schema

Prompt Structure (Subtask 1: Flu Vaccination Status):

```
// SYSTEM PROMPT
You are an expert annotator for a health informatics
     shared task. Your job is to classify social
     media posts (tweets) according to the author's
     self-reported flu vaccination status.

Classify each tweet into exactly one of the
     following five categories:
```

```
- Currently-Vaccinated: The author describes a
    SPECIFIC, COMPLETED vaccination event in the
    current or most recent flu season. Look for
    past-tense or just-completed language: "I got
    my flu shot", "just got vaccinated", "got my
    flu shot today", "I received my flu vaccine
    last week". The key requirement is that the
    author confirms the shot already happened
    during this season.

- Currently-Unvaccinated: The author explicitly or
    implicitly indicates they have NOT received and
     do not intend to receive a flu shot. This
    includes:
  - Direct statements: "I don't get a flu shot", "I'
      m not getting the flu vaccine"
  - Habitual negatives: "I've never gotten a flu
      shot in my life", "I never get flu shots"
  - Dismissals or refusals: "no thanks", "I'll pass
      ", "so I'm good" (in context of declining)
  - Past vaccination followed by explicit cessation:
       "I got the flu shot years ago but stopped
      getting them", "Got one once, never again"
  This category takes priority over "Previously-
      Vaccinated" when the author describes past
      vaccination but also makes clear they no
      longer get vaccinated.

- Previously-Vaccinated: The author mentions
    receiving a flu shot in a clearly prior flu
    season (e.g., "I got a flu shot last year", "
    back in 2019 I got vaccinated") with no
    indication of their current season status AND
    no indication they stopped getting vaccinated.
    If the author also expresses that they will not
     get one this year or have stopped, categorize
    as "Currently-Unvaccinated".

- Possibly-Vaccinated: The author expresses intent,
    consideration, plans, or obligation to get a
    flu shot but does not confirm they have
    actually received one. This includes:
  - Intent/plans: "thinking about getting a flu shot
      ", "I need to get my flu shot", "getting my
      flu shot this week"
  - Habitual pattern (no specific event): "I get the
       flu shot every year", "I always get a flu
      shot", "I take the flu vaccine annually"
  - Obligation/mandate: "We have to get a flu shot
      for work", "I'm required to get the flu
      vaccine", "why do I have to take a flu
      vaccine"
  The distinction from Currently-Vaccinated is
      critical: habitual present-tense statements
      ("I get my flu shot every year") describe a
      PATTERN, not a specific completed event.
      Obligation statements describe a requirement,
       not a completed event. Both should be
      classified as Possibly-Vaccinated unless the
      tweet also describes a specific current-
      season vaccination event.

- Other: The tweet mentions flu vaccination but does
     NOT describe the author's own personal
    vaccination status. This includes: discussing
    someone else's vaccination, general advocacy or
     encouragement to vaccinate, sharing news or
    statistics, rhetorical or hypothetical
    statements, or any tweet where the author's own
     status cannot be determined.

Decision rules:
1. The classification must be based solely on what
    the author explicitly states about themselves -
     do not infer unstated information.
2. "Currently" refers to the present or most recent
    flu season. If the timing is ambiguous but the
    language is present-tense or recent, treat it
    as current.
3. If the tweet is PRIMARILY about a third party (
    child, family member, public figures) or is
    general commentary, classify as Other - UNLESS
    the author also reveals their own vaccination
    status in the tweet, in which case classify
    based on the author's status.
4. Habitual negatives ("I've never gotten one", "I
    never get flu shots") imply ongoing current
    status and should be classified as Currently-
    Unvaccinated.
5. If the tweet contains both past vaccination and
    current uncertainty, use the most specific
    information available.
6. CRITICAL - Habitual present tense vs. completed
    event:
  - "I get my flu shot every year" -> Possibly-
      Vaccinated (habitual pattern, no specific
      event confirmed)
  - "I got my flu shot" / "I got my flu shot today"
       -> Currently-Vaccinated (specific completed
       event)
  - "I'm getting my flu shot tomorrow" -> Possibly-
      Vaccinated (planned, not yet completed)
  - "Wore a mask to get a flu shot" -> Currently-
      Vaccinated (implies a specific completed
      visit)
7. Temporal references like "a few months ago", "in
    October", "last month" WITHOUT a prior-year
    reference should be treated as current season,
    not previously vaccinated.
8. Past vaccination + explicit cessation = Currently
    -Unvaccinated, not Previously-Vaccinated:
  - "I got the flu shot but stopped after that" ->
      Currently-Unvaccinated
  - "Got a flu vaccine once... never got another"
      -> Currently-Unvaccinated
9. Implicit refusals or dismissals count as
    Currently-Unvaccinated:
  - "so I'm good" (in context of not wanting a flu
      shot) -> Currently-Unvaccinated
  - "no thanks" / "I'll pass on the flu shot" ->
      Currently-Unvaccinated
10. Obligation/mandate language about the author
    implies Possibly-Vaccinated (intent through
    requirement), not Other:
   - "I have to get a flu shot for work" ->
       Possibly-Vaccinated
   - "We're required to get the flu vaccine" ->
       Possibly-Vaccinated

First reason through the tweet carefully, then
    provide your label.

// USER PROMPT
Tweet Created At: {timestamp}
Tweet: {raw_tweet_text}
```

JSON Schema (Subtask 1: Flu Vaccination Status Classification):

```
{
  "type": "json_schema",
  "name": "flu_vaccination_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning
            explaining which evidence in the tweet
            supports the chosen label, and why
            other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Vaccinated",
          "Currently-Unvaccinated",
          "Previously-Vaccinated",
          "Possibly-Vaccinated",
          "Other"
        ],
        "description": "The flu vaccination status
            classification label."
      }
    },
    "required": [
      "reasoning",
      "label"
    ],
    "additionalProperties": false
  }
```

```
}
```

Prompt Structure (Subtask 2: Flu Test Result):

```
// SYSTEM PROMPT
You are an expert annotator for a health informatics
     shared task. Your job is to classify social
    media posts (tweets) according to the author's
    self-reported flu test result or flu diagnosis.

Classify each tweet into exactly one of the
    following five categories:

- Currently-Positive: The author explicitly reports
    a recent positive flu test result or a current
    flu diagnosis (e.g., "I tested positive for flu
    ", "just diagnosed with the flu", "I have the
    flu"). The result pertains to the current or
    very recent illness episode.

- Currently-Negative: The author explicitly reports
    a recent negative flu test result or a current
    diagnosis that excludes flu (e.g., "my flu test
     came back negative", "tested negative for flu
    ", "they ruled out the flu"). The result
    pertains to the current or very recent illness
    episode.

- Previously-Positive: The author mentions having
    tested positive for flu or been diagnosed with
    flu during a clearly past season or episode,
    with no indication of a current test (e.g., "I
    had flu last January", "I was positive for flu
    back in 2019").

- Previously-Negative: The author mentions having
    tested negative for flu or been told they did
    not have flu during a clearly past episode,
    with no indication of a current test (e.g., "
    they tested me last year and flu came back
    negative").

- Other: The tweet mentions flu testing or diagnosis
     but does NOT describe the author's own
    personal test result. This includes: discussing
     someone else's test, general commentary about
    flu testing practices, encouraging others to
    get tested, reporting statistics, ambiguous or
    unresolvable statements, or any tweet where the
     author's own test result cannot be determined.

Decision rules:
1. The classification must be based solely on what
    the author explicitly states about their own
    flu test or diagnosis - do not infer unstated
    results.
2. "Currently" refers to a recent, ongoing, or just-
    concluded illness episode. Past-tense language
    describing events in a prior season should be
    classified as Previously-Positive or Previously
    -Negative.
3. If the tweet discusses someone else's test result
     (child, family, public figures) or is general
    commentary, classify as Other.
4. Symptoms alone (fever, cough) without a confirmed
     test result or diagnosis do not constitute a
    test result - classify as Other.
5. If multiple flu tests are mentioned (current and
    past), classify based on the most recent one.
6. A positive COVID test or other non-flu diagnosis
    does not qualify as a flu test result -
    classify as Other unless flu is also explicitly
     mentioned.

First reason through the tweet carefully, then
    provide your label.

// USER PROMPT
Tweet Created At: {timestamp}
Tweet: {raw_tweet_text}
```

JSON Schema (Subtask 2: Flu Test Result Classification):

```
{
  "type": "json_schema",
  "name": "flu_test_result_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning
            explaining which evidence in the tweet
            supports the chosen label, and why
            other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Positive",
          "Currently-Negative",
          "Previously-Positive",
          "Previously-Negative",
          "Other"
        ],
        "description": "The flu test result
            classification label."
      }
    },
    "required": [
      "reasoning",
      "label"
    ],
    "additionalProperties": false
  }
}
```

JSON Schema (Subtask 2: Flu Test Result Classification):

```
{
  "type": "json_schema",
  "name": "flu_test_result_classification",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "reasoning": {
        "type": "string",
        "description": "Step-by-step reasoning
            explaining which evidence in the tweet
            supports the chosen label, and why
            other labels were ruled out."
      },
      "label": {
        "type": "string",
        "enum": [
          "Currently-Positive",
          "Currently-Negative",
          "Previously-Positive",
          "Previously-Negative",
          "Other"
        ],
        "description": "The flu test result
            classification label."
      }
    },
    "required": [
      "reasoning",
      "label"
    ],
    "additionalProperties": false
  }
}
```

A.6 Task 4: Dial2Note Prompts and Schema

Prompt Structure:

```
// SYSTEM PROMPT
You are a clinical documentation specialist. Given a
     doctor-patient dialogue, generate a structured
     SOAP note (Subjective, Objective, Assessment,
    Plan).
```

```
GROUNDING RULE - every statement in the note must be
      directly supported by the dialogue:
- Demographics (name, age, sex, ethnicity): include
     ONLY if explicitly stated. Do not infer from
     names or voice.
- Doctor/specialist names: NEVER fabricate. Write "
     Refer to [specialty]" unless the doctor names
     them.
- Diagnoses: use only diagnoses the doctor
     explicitly states.

Instructions:
- Subjective: Document CC, HPI, ROS, PMH, surgical/
     family/social history, allergies, and
     medications. Keep the HPI concise - summarize
     in 2-4 sentences, don't restate every exchange.
      Include severity in the CC (e.g. "Moderate to
     severe shortness of breath").
- Objective: Record vital signs, physical exam, and
     test results. Use the doctor's exact clinical
     language for exam findings. When a vital is "
     normal" without a number, write "Normal (
     specific value not provided)".
- Assessment: Use a bulleted list of diagnoses.
     Include the doctor's clinical reasoning when
     stated. Include differential diagnoses if the
     doctor discusses them. Include ICD-10 codes if
     the doctor mentions them.
- Plan: Group items under bold sub-headings. Common
     headings: **Medical Treatment:**, **Medications
     :**, **Imaging:**, **Investigations:**, **
     Referrals:**, **Lifestyle Modifications:**, **
     Follow-up:**, **Patient Education and
     Counseling:**, **Patient Agreements:**. Under
     each heading, list items as indented sub-
     bullets. Use the doctor's own words. Only
     include items explicitly discussed.
- When the patient explicitly agrees with the plan
     or the doctor confirms understanding, include
     **Patient Agreements:** as a final plan block.

Formatting rules:
- Use null for fields not mentioned; empty arrays []
      for empty lists.
- For ROS and physical exam entries, use only
     canonical system names from the enum.
- Write current_medications as prose strings (e.g. "
     Metformin 500 mg twice daily for diabetes").
- Each plan item in the JSON array should be a
     category block: bold heading (e.g. "**
     Medications:**") followed by newline-indented
     sub-items ("\n  - Clindamycin 300 mg orally
     every 6 hours.").
- Social history = patient's reported habits, NOT
     doctor's advice.
- ROS = only systems the doctor asked about or the
     patient reported symptoms for.
- Be concise but complete.

Below are three example dialogues and their expected
      SOAP notes. Note how each note includes ONLY
     information grounded in the dialogue.

--- EXAMPLE 1 (Rheumatology - Joint Pain) ---

Dialogue:
[doctor]: Hi, how are you doing today? What brings
     you in?
[patient]: Hi, doctor. I've been having some pretty
     bad joint pain, stiffness, and swelling lately.
[doctor]: I'm sorry to hear that. Can you tell me a
     bit more about these symptoms? How long have
     you been experiencing them?
[patient]: It's been about 6 months now. The pain is
      there every day and it's worse in the mornings
     . It's really affecting my ability to type and
     use a mouse at work.
[doctor]: I see. That sounds quite uncomfortable. Is
      there any redness or restricted range of
     motion in your joints?
[patient]: Yes, occasionally there's redness and my
     movement is definitely restricted. It's made my
      daily exercise routine almost impossible.
[doctor]: That sounds challenging. Have you noticed
     any other symptoms like fever or weight loss?
[patient]: No, no fever or weight loss.
[doctor]: Hmm, okay. How about your cardiovascular
     health? Any known issues there?
[patient]: I do have hypertension, but it's
     currently managed with medication.
[doctor]: Alright. And any skin issues?
[patient]: Yes, I have psoriasis. I'm managing it
     with topical corticosteroids.
[doctor]: Thank you for sharing that. Let's go over
     your vital signs. Your blood pressure is 130/85
      mmHg, heart rate is 70 bpm, respiratory rate
     is 16 breaths per minute, and your temperature
     is 98.6 degrees F.
[patient]: Sounds normal, I guess?
[doctor]: Yes, those are within normal ranges. Now,
     let's move on to the physical examination. I'll
      check your joints and skin.
[patient]: Okay.
[doctor]: I see tenderness and swelling in your
     wrists, knees, and fingers. There's reduced
     range of motion in these joints, but no
     deformities or subluxations. Your skin has
     scaly plaques consistent with psoriasis.
[patient]: That matches what I've been feeling.
[doctor]: We'll need to do some investigations,
     including X-rays of the affected joints and
     blood work-Complete Blood Count (CBC), C-
     Reactive Protein (CRP), Erythrocyte
     Sedimentation Rate (ESR), and Rheumatoid Factor
      (RF).
[patient]: Alright.
[doctor]: Based on your symptoms and history, the
     primary diagnosis is Psoriatic Arthropathy. We
     also need to consider your psoriasis and
     hypertension.
[patient]: Okay, what's the plan?
[doctor]: For pain relief, I'll prescribe Naproxen
     500 mg, to be taken orally twice daily. To slow
      the progression of your joint issues, we'll
     start Methotrexate 10 mg, taken orally once a
     week. You'll also need to take Folic acid 1 mg
     daily to counteract potential side effects of
     Methotrexate.
[patient]: Got it. Anything else?
[doctor]: I'll refer you to a physical therapist, Dr
     . Allison Smith, for joint mobility exercises
     and pain management techniques.
[patient]: That sounds helpful.
[doctor]: We'll need to schedule a follow-up
     appointment in 4 weeks to monitor how the
     medications are working and check for any side
     effects. Also, we'll order liver function tests
      every 6 weeks to ensure Methotrexate is safe
     for you.
[patient]: Okay, I understand.
[doctor]: Do you have any questions or concerns
     right now?
[patient]: Not at the moment. Just want to get this
     under control.
[doctor]: That's completely understandable. Make
     sure to follow up in 4 weeks and contact the
     clinic if you experience any adverse effects or
      worsening of symptoms.
[patient]: Will do. Thanks, doctor.
[doctor]: You're welcome. Take care and see you in 4
      weeks.
[patient]: Thanks, you too.

Expected note:
{
  "subjective": {
    "chief_complaint": "Joint pain, stiffness, and
         swelling for 6 months",
    "hpi": "The patient presents with joint pain,
         stiffness, and swelling for 6 months. The
         pain is daily and worse in the mornings,
         impacting ability to type and use a mouse
         at work. Occasional redness and restricted
         range of motion noted, hindering daily
         exercise.",
    "ros": [
      {"system": "Musculoskeletal", "finding": "
           Positive for joint pain, swelling,
           stiffness, occasional redness, and
           restricted range of motion"},
      {"system": "Constitutional", "finding": "
           Denies fever, weight loss"},
```

```
      {"system": "Cardiovascular", "finding": "Known
           history of hypertension, currently
          managed with medication"},
      {"system": "Skin", "finding": "Positive for
           psoriasis, currently managed with topical
           corticosteroids"}
    ],
    "past_medical_history": ["Hypertension", "
        Psoriasis"],
    "surgical_history": [],
    "family_history": [],
    "social_history": null,
    "allergies": [],
    "current_medications": ["Topical corticosteroids
         for psoriasis", "Antihypertensive
        medication (unspecified)"]
  },
  "objective": {
    "vital_signs": {
      "blood_pressure": "130/85 mmHg",
      "heart_rate": "70 bpm",
      "respiratory_rate": "16 breaths per minute",
      "temperature": "98.6 F",
      "oxygen_saturation": null,
      "weight": null, "height": null, "bmi": null, "
           pain_scale": null
    },
    "physical_exam": [
      {"system": "Musculoskeletal", "finding": "
           Tenderness and swelling observed in
           multiple joints including wrists, knees,
           and fingers. Reduced range of motion
           noted in the affected joints. No
           deformities or subluxations detected."},
      {"system": "Skin", "finding": "Scaly plaques
           consistent with psoriasis"}
    ],
    "test_results": []
  },
  "assessment": "- Psoriatic Arthropathy\n-
      Psoriasis\n- Hypertension",
  "plan": [
    "**Medical Treatment:**\n - Prescribe Naproxen
        500 mg, oral, twice daily for pain relief.\
        n - Start Methotrexate 10 mg, oral, once
        weekly to slow disease progression.\n -
        Advise taking Folic acid 1 mg, oral, daily
        to mitigate Methotrexate side effects.",
    "**Investigations:**\n - Order X-ray of
        affected joints (wrists, knees, fingers).\n
         - Order blood work: Complete Blood Count
        (CBC), C-Reactive Protein (CRP),
        Erythrocyte Sedimentation Rate (ESR),
        Rheumatoid Factor (RF).",
    "**Referrals:**\n - Refer to physical therapist
         for joint mobility exercises and pain
        management techniques.",
    "**Follow-up:**\n - Schedule follow-up
        appointment in 4 weeks to monitor
        medication efficacy and any potential side
        effects.\n - Order liver function tests
        every 6 weeks to ensure the safety of
        Methotrexate.",
    "**Patient Agreements:**\n - The patient
        understands and agrees with the recommended
         medical treatment plan.\n - Follow up in
        4 weeks for evaluation.\n - Contact the
        clinic if experiencing any adverse effects
        or worsening of symptoms."
  ]
}

--- EXAMPLE 2 (Obstetrics - Labor with Meconium) ---

Dialogue:
[doctor]: Good morning, how are you feeling today?
[patient]: Good morning, Doctor. I'm feeling a bit
    anxious, to be honest.
[doctor]: I understand. Can you tell me a bit more
    about what brought you in today?
[patient]: Well, I'm in labor and my water broke.
    The fluid was a greenish color, and I got
    really worried.
[doctor]: I see. How long ago did your labor start?
[patient]: About 12 hours ago. I had some clear and
    bloody show before my water broke.
[doctor]: Okay, thank you for sharing that. Have you
     experienced any fever or chills?
[patient]: No, none at all.
[doctor]: How about nausea or vomiting during labor?
[patient]: No, I haven't had any nausea or vomiting.
[doctor]: Any shortness of breath?
[patient]: No, my breathing has been normal.
[doctor]: And how about urination? Any burning or
    discomfort?
[patient]: No, my urination has been normal.
[doctor]: Any significant pain aside from the labor
    contractions?
[patient]: No, just the usual labor pains.
[doctor]: Alright, let's review your vital signs.
    Your blood pressure is 120/80 mmHg, heart rate
    is 85 bpm, temperature is 98.6 degrees F, and
    respiratory rate is 18 breaths/min. All stable.
[patient]: That's good to hear.
[doctor]: On physical examination, your fundal
    height is consistent with term pregnancy and
    the baby is in a vertex position. Your cervix
    is fully dilated and 100% effaced. We did
    notice decelerations on the continuous
    electronic fetal monitoring, but we managed
    them with maternal positional changes and
    supplemental oxygen.
[patient]: Okay, that sounds a bit concerning.
[doctor]: Yes, and the greenish discoloration of the
     amniotic fluid indicates meconium staining,
    which suggests fetal stress.
[patient]: What does that mean for the baby?
[doctor]: The presence of meconium in the amniotic
    fluid can potentially lead to aspiration and
    respiratory complications for the newborn. We
    will need to take immediate action once the
    baby is born.
[patient]: What kind of actions?
[doctor]: We will need to suction the infant's
    airway immediately at birth to clear any
    meconium that might be present. The newborn
    will also be admitted to the Neonatal Intensive
     Care Unit (NICU) for close monitoring of their
     respiratory status.
[patient]: That sounds serious. Will the baby need
    oxygen?
[doctor]: If necessary, we will provide supplemental
     oxygen to manage any respiratory distress. We
    will also refer to Neonatology for
    comprehensive evaluation and management.
[patient]: Is there anything I need to do?
[doctor]: After delivery, we will have a pediatric
    follow-up within one week post-discharge to
    monitor for any delayed respiratory symptoms or
     other complications. We will also continue to
    monitor and manage your gestational diabetes
    through the postpartum period.
[patient]: Okay, what should I look out for in the
    baby?
[doctor]: It's important to recognize signs of
    respiratory distress in the newborn, such as
    cyanosis, rapid breathing, or grunting. I will
    discuss this in more detail with you and your
    partner.
[patient]: Alright, I'll keep an eye out.
[doctor]: We will also discuss maintaining a
    balanced diet and monitoring blood glucose
    levels post-delivery. Mild exercise routines
    postpartum can aid in recovery, but we will
    provide specific advice based on your condition
    .
[patient]: That sounds good.
[doctor]: Do you have any questions or concerns?
[patient]: No, I think you've covered everything.
    Thank you for explaining it all.
[doctor]: You're welcome. Our priority is the health
     and safety of both you and your baby. We'll be
     here every step of the way. Let's get you
    ready for delivery.
[patient]: Thank you, Doctor. I appreciate it.
[doctor]: You're welcome. Let's proceed and ensure
    everything goes smoothly.

Expected note:
{
  "subjective": {
    "chief_complaint": "Labor and delivery
         complicated by meconium in amniotic fluid",
```

```
  "hpi": "The patient is presenting with labor at
       term. She noticed rupture of membranes
       accompanied by greenish discoloration of
       the amniotic fluid, indicative of meconium
       staining. Labor onset was 12 hours ago with
        clear and bloody show prior to membrane
       rupture.",
  "ros": [
    {"system": "Constitutional", "finding": "No
       fever or chills"},
    {"system": "Gastrointestinal", "finding": "No
       nausea or vomiting during labor"},
    {"system": "Respiratory", "finding": "No
       shortness of breath"},
    {"system": "Genitourinary", "finding": "Normal
        urination, no dysuria"},
    {"system": "Musculoskeletal", "finding": "No
       significant pain other than labor
       contractions"}
  ],
  "past_medical_history": ["Gestational diabetes
      "],
  "surgical_history": [],
  "family_history": [],
  "social_history": null,
  "allergies": [],
  "current_medications": []
},
"objective": {
  "vital_signs": {
    "blood_pressure": "120/80 mmHg",
    "heart_rate": "85 bpm",
    "respiratory_rate": "18 breaths/min",
    "temperature": "98.6 F",
    "oxygen_saturation": null,
    "weight": null, "height": null, "bmi": null, "
       pain_scale": null
  },
  "physical_exam": [
    {"system": "Abdomen", "finding": "Fundal
       height consistent with term pregnancy,
       fetal position vertex"},
    {"system": "Pelvic", "finding": "Cervix fully
       dilated, 100% effaced"}
  ],
  "test_results": ["Continuous electronic fetal
       monitoring: decelerations noted, managed
       with maternal positional changes and
       supplemental oxygen", "Amniotic fluid:
       positive for meconium staining"]
},
"assessment": "- Labor and delivery complicated by
     meconium in amniotic fluid.\n- Medical
     Reasoning: The presence of meconium in the
     amniotic fluid suggests fetal stress, which
     could lead to aspiration and potential
     respiratory complications for the newborn.",
"plan": [
  "**Medical Treatment:**\n - Immediate
       suctioning of the infant's airway at birth
       to clear any meconium present.\n -
       Admission of the newborn to the Neonatal
       Intensive Care Unit (NICU) for close
       monitoring of respiratory status.\n -
       Supplemental oxygen to the newborn if
       necessary to manage any respiratory
       distress.",
  "**Referrals:**\n - Referral to Neonatology for
        comprehensive evaluation and management of
        potential aspiration and respiratory
       complications.",
  "**Follow-up:**\n - Pediatric follow-up within
       1 week post-discharge to monitor for
       delayed respiratory symptoms or other
       complications.\n - Continued monitoring
       and management of gestational diabetes
       through the postpartum period.",
  "**Patient Education and Counseling:**\n -
       Discussed with the patient the importance
       of recognizing signs of respiratory
       distress in the newborn, such as cyanosis,
       rapid breathing, or grunting.\n -
       Discussed maintaining a balanced diet and
       monitoring blood glucose levels post-
       delivery.\n - Encouraged mild exercise
       routines postpartum to aid in recovery.",
    "**Patient Agreements:**\n - The patient
       understands and agrees with the recommended
        medical treatment plan."
  ]
}
```

--- EXAMPLE 3 (Emergency - Sports Injury) ---

Dialogue:
[doctor]: Hi there, I understand you're experiencing some pain and swelling in your right lower leg. Can you tell me what happened?
[patient]: Yes, I was playing in a local sports league game about two hours ago and I hurt my leg. It's been hurting and swelling up since then.
[doctor]: Oh, that sounds painful. Can you describe the pain for me? How severe is it on a scale of 1 to 10?
[patient]: It's pretty bad, I'd say it's around a 6 or 7. It hurts a lot, and I can't walk or put any weight on it.
[doctor]: I see. Did you notice any bruising or other symptoms?
[patient]: Yes, there's some slight bruising and it's swollen. The pain has been continuous since the injury.
[doctor]: Alright, let's go over some other symptoms. Have you had any fever or chills?
[patient]: No, none of that.
[doctor]: How about chest pain or palpitations?
[patient]: No, my chest feels fine.
[doctor]: Any shortness of breath or cough?
[patient]: No, nothing like that.
[doctor]: Aside from the pain and swelling in your right lower leg, are you experiencing any other joint pain or muscle weakness?
[patient]: No, just the leg.
[doctor]: Okay, thank you for that information. Let's take a look at your vital signs. Your blood pressure is 122 over 78, heart rate is 78 beats per minute, respiratory rate is 16 breaths per minute, temperature is 98.6 degrees Fahrenheit, and your oxygen saturation is 98% on room air.
[patient]: Okay, that sounds good, I guess.
[doctor]: Now, let's do a physical examination. I'll start with listening to your heart and lungs. Your cardiovascular system shows a regular rate and rhythm with no murmurs. Your lungs are clear to auscultation bilaterally.
[patient]: Alright.
[doctor]: I'm going to examine your right lower leg now. I see there's swelling and tenderness. How does it feel when I press here?
[patient]: Ouch, that hurts.
[doctor]: Your range of motion is limited due to pain. I don't see any obvious deformities, and there's no crepitus or instability noted. We should get an X-ray to rule out any fractures or other significant injuries. Does that sound okay?
[patient]: Yes, that's fine.
[doctor]: While we wait for the X-ray results, let's discuss the plan. I recommend applying ice to your leg every 2 hours for the next 48 hours to reduce swelling. Elevate your leg above heart level as much as possible.
[patient]: Alright, I can do that.
[doctor]: Also, avoid putting weight on your leg. I'll provide you with crutches. For the pain, I'll prescribe ibuprofen 400 mg tablets. Take one every 6 hours as needed, up to 30 tablets.
[patient]: Got it, thanks.
[doctor]: Make sure to follow up with your primary care physician in one week for reassessment. If you experience increased pain, swelling, redness, or if you cannot move your leg, return to the emergency department immediately.
[patient]: Okay, I'll keep an eye on it.
[doctor]: Do you have any questions about the instructions?
[patient]: No, I think I got it. Thank you.
[doctor]: Great. We'll wait for the X-ray results and then proceed accordingly. Take care and rest that leg.
[patient]: Thank you, Doctor.

```
[doctor]: You're welcome. I'll be back shortly once
    we have the X-ray results.

Expected note:
{
  "subjective": {
    "chief_complaint": "Moderate pain and swelling
        in the right lower leg",
    "hpi": "The patient presents with pain and
        swelling in the right lower leg. The injury
         occurred while participating in a local
        sports league game about 2 hours ago.
        Reports continuous pain since the injury,
        accompanied by swelling and slight bruising
        . Pain severity is 6-7/10. Unable to walk
        or bear weight on the affected leg.",
    "ros": [
      {"system": "Constitutional", "finding": "
          Denies fever, chills"},
      {"system": "Cardiovascular", "finding": "
          Denies chest pain, palpitations"},
      {"system": "Respiratory", "finding": "Denies
          shortness of breath, cough"},
      {"system": "Musculoskeletal", "finding": "
          Reports pain, swelling, and slight
          bruising in the right lower leg; denies
          other joint pain or muscle weakness"}
    ],
    "past_medical_history": [],
    "surgical_history": [],
    "family_history": [],
    "social_history": null,
    "allergies": [],
    "current_medications": []
  },
  "objective": {
    "vital_signs": {
      "blood_pressure": "122/78 mmHg",
      "heart_rate": "78 bpm",
      "respiratory_rate": "16 breaths per minute",
      "temperature": "98.6 F",
      "oxygen_saturation": "98% on room air",
      "weight": null, "height": null, "bmi": null, "
          pain_scale": "6-7/10"
    },
    "physical_exam": [
      {"system": "Cardiovascular", "finding": "
          Regular rate and rhythm, no murmurs"},
      {"system": "Respiratory", "finding": "Lungs
          clear to auscultation bilaterally"},
      {"system": "Extremities", "finding": "Right
          lower leg: swelling and tenderness noted,
           limited range of motion due to pain, no
          obvious deformities, no crepitus or
          instability noted"}
    ],
    "test_results": ["X-ray of right lower leg
        ordered to rule out fractures or other
        significant injuries; results pending"]
  },
  "assessment": "- Right lower leg injury with pain,
       swelling, and inability to bear weight.\n-
      Differential: Possible fracture vs soft
      tissue injury; X-ray ordered.",
  "plan": [
    "**Imaging:**\n  - X-ray of right lower leg to
        rule out fractures or other significant
        injuries. Await and review results.",
    "**Medical Treatment:**\n  - Ibuprofen 400 mg
        tablets, oral, every 6 hours as needed for
        pain; 30 tablets prescribed.\n  - Apply ice
         to leg every 2 hours for the next 48 hours
         to reduce swelling.\n  - Elevate leg above
         heart level as much as possible.\n  -
        Avoid weight-bearing on the affected leg;
        crutches provided.",
    "**Follow-up Recommendations:**\n  - Follow up
        with primary care physician in 1 week for
        reassessment.",
    "**Patient Education and Counseling:**\n  -
        Patient advised to rest and restrict
        activities that could exacerbate the injury
        .\n  - Return to the emergency department
        immediately for increased pain, swelling,
        redness, or inability to move the leg."
  ]
}
```

```
// USER PROMPT
Dialogue: {raw_dialogue_text}
```

JSON Schema:

```
{
  "$schema": "http://json-schema.org/draft-07/schema
      #",
  "title": "SOAPNoteProse",
  "description": "SOAP note with prose-style fields
      optimised for lexical overlap with MedSynth
      reference notes",
  "type": "object",
  "additionalProperties": false,
  "required": [
    "subjective",
    "objective",
    "assessment",
    "plan"
  ],
  "properties": {
    "subjective": {
      "$ref": "#/$defs/Subjective"
    },
    "objective": {
      "$ref": "#/$defs/Objective"
    },
    "assessment": {
      "type": "string",
      "description": "Clinical assessment. May be a
          single diagnosis line, a bulleted list of
           diagnoses, or a short narrative
          paragraph synthesising the findings -
          match whichever style the clinical
          context warrants. Include ICD-10 codes
          only if explicitly mentioned in the
          dialogue."
    },
    "plan": {
      "type": "array",
      "description": "Ordered list of plan items
          written in clinical prose. Each item
          should be a complete sentence or short
          paragraph. Prefix with a category label
          when helpful (e.g. 'Medication:', '
          Imaging:', 'Referral:', 'Follow-up:').
          Mirror the doctor's exact wording from
          the dialogue wherever possible.",
      "items": {
        "type": "string"
      }
    }
  },
  "$defs": {
    "Subjective": {
      "type": "object",
      "additionalProperties": false,
      "required": [
        "chief_complaint",
        "hpi",
        "ros",
        "past_medical_history",
        "surgical_history",
        "family_history",
        "social_history",
        "allergies",
        "current_medications"
      ],
      "properties": {
        "chief_complaint": {
          "type": [
            "string",
            "null"
          ],
          "description": "Brief statement of the
              patient's presenting problem"
        },
        "hpi": {
          "type": [
            "string",
            "null"
          ],
          "description": "History of Present Illness
               - narrative paragraph. Always
              include the patient's full name, age,
              sex, and ethnicity when available in
              the dialogue."
```

```
    },
    "ros": {
      "type": "array",
      "description": "Review of Systems entries
          ",
      "items": {
        "$ref": "#/$defs/ROSEntry"
      }
    },
    "past_medical_history": {
      "type": "array",
      "description": "List of pertinent current
          or past medical conditions",
      "items": {
        "type": "string"
      }
    },
    "surgical_history": {
      "type": "array",
      "description": "Prior surgeries as prose
          strings, e.g. 'Appendectomy (2018)'",
      "items": {
        "type": "string"
      }
    },
    "family_history": {
      "type": "array",
      "description": "Family history as prose
          strings, e.g. 'Father: hypertension,
          diabetes'",
      "items": {
        "type": "string"
      }
    },
    "social_history": {
      "type": [
        "string",
        "null"
      ],
      "description": "Social history as a short
          prose paragraph or semicolon-
          separated summary covering tobacco,
          alcohol, drug use, diet, exercise,
          occupation, etc. Use null if not
          discussed."
    },
    "allergies": {
      "type": "array",
      "description": "Free-text allergy strings,
          e.g. 'Penicillin - rash'",
      "items": {
        "type": "string"
      }
    },
    "current_medications": {
      "type": "array",
      "description": "Current medications as
          prose strings mirroring the dialogue
          wording, e.g. 'Metformin 500 mg twice
           daily for diabetes'",
      "items": {
        "type": "string"
      }
    }
  }
},
"Objective": {
  "type": "object",
  "additionalProperties": false,
  "required": [
    "vital_signs",
    "physical_exam",
    "test_results"
  ],
  "properties": {
    "vital_signs": {
      "anyOf": [
        {
          "$ref": "#/$defs/VitalSigns"
        },
        {
          "type": "null"
        }
      ]
    },
    "physical_exam": {
      "type": "array",
      "description": "Physical exam findings by
          body system or region",
      "items": {
        "$ref": "#/$defs/PhysicalExamEntry"
      }
    },
    "test_results": {
      "type": "array",
      "description": "Completed test results as
          prose strings, e.g. 'Chest X-ray: No
          acute findings'",
      "items": {
        "type": "string"
      }
    }
  }
},
"ROSEntry": {
  "type": "object",
  "additionalProperties": false,
  "required": [
    "system",
    "finding"
  ],
  "properties": {
    "system": {
      "type": "string",
      "enum": [
        "Constitutional",
        "Cardiovascular",
        "Respiratory",
        "Gastrointestinal",
        "Genitourinary",
        "Musculoskeletal",
        "Neurological",
        "Psychiatric",
        "Endocrine",
        "Skin",
        "HEENT",
        "Eyes",
        "ENT",
        "Hematologic",
        "Allergic",
        "Breast",
        "Gynecological",
        "Sleep",
        "Other"
      ]
    },
    "finding": {
      "type": "string",
      "description": "Clinical finding or
          symptom report for this system"
    }
  }
},
"PhysicalExamEntry": {
  "type": "object",
  "additionalProperties": false,
  "required": [
    "system",
    "finding"
  ],
  "properties": {
    "system": {
      "type": "string",
      "enum": [
        "General",
        "HEENT",
        "Neck",
        "Cardiovascular",
        "Respiratory",
        "Abdomen",
        "Musculoskeletal",
        "Neurological",
        "Skin",
        "Extremities",
        "Psychiatric",
        "Genitourinary",
        "Pelvic",
        "Rectal",
        "Breast",
        "Eyes"
      ]
    },
    "finding": {
      "type": "string",
```

```
        "description": "Objective examination
            finding for this region"
      }
    }
  },
  "VitalSigns": {
    "description": "All values stored as strings
        to preserve units and qualifiers",
    "type": "object",
    "additionalProperties": false,
    "required": [
      "blood_pressure",
      "heart_rate",
      "respiratory_rate",
      "temperature",
      "oxygen_saturation",
      "weight",
      "height",
      "bmi",
      "pain_scale"
    ],
    "properties": {
      "blood_pressure": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '120/80 mmHg'"
      },
      "heart_rate": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '72 bpm'"
      },
      "respiratory_rate": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '16 breaths/min'"
      },
      "temperature": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '98.6 degrees F'"
      },
      "oxygen_saturation": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '98% on room air'"
      },
      "weight": {
        "type": [
          "string",
          "null"
        ]
      },
      "height": {
        "type": [
          "string",
          "null"
        ]
      },
      "bmi": {
        "type": [
          "string",
          "null"
        ]
      },
      "pain_scale": {
        "type": [
          "string",
          "null"
        ],
        "description": "e.g. '4/10'"
      }
    }
  }
}
```

A.7 Task 5: Patient Metadata Prompts and Schema

Prompt Structure:

```
// SYSTEM PROMPT
You are an expert biomedical NLP system and
    epidemiological data auditor. Your task is to
    analyze sentences extracted from PubMed
    articles that report generating SARS-CoV-2
    sequences, and determine if the sentence
    contains a valid mention of Patient Metadata
    for the study's own patients or samples.

DEFINITION OF PATIENT METADATA CATEGORIES:
"age", "sex", "race/ethnicity", "symptoms", "
    disease_severity", "viral_load", "
    duration_of_infection", "lab_results", "
    vital_signs", "treatments", "hospitalization",
    "outcomes", "comorbidities", "risk_factors", "
    vaccination_status", "place_of_residence", "
    geographic_location", "travel_history".

WHAT COUNTS AS "geographic_location" (this category
    is commonly misclassified):
POSITIVE - any of these patterns:
- The sentence names WHERE the study's own samples,
    sequences, or patients are from: "Indian SARS-
    CoV-2 sequences", "44 Hong Kong strains", "
    Maradi genomes", "Phylogenetic Analysis of SARS
    -CoV-2 in Colombia" (section title naming the
    study country)
- The sentence describes where samples were
    collected or received: "Nasopharyngeal swabs
    were received from the Civil Hospital Hisar,
    Haryana (India)"
- The sentence attributes sequences to a specific
    location in the current study: "residence-hall
    sequences clustered with viruses circulating in
     Dane County"
- The sentence reports testing or positivity rates
    for the study's geographic area: "10.8% overall
     positivity rate for the Central Texas
    geography"
NEGATIVE - any of these patterns:
- Regional case counts or epidemiological statistics
     not tied to specific patients: "307 infections
     in New South Wales"
- Where a variant/lineage was first globally
    detected: "B.1.1.7 was first detected in the UK
    "
- Phylogenetic comparison to EXTERNAL reference
    sequences from other countries: "more closely
    related to strains seen in South Korea (n = 7)
    and France (n = 5)" when South Korea and France
     are NOT the study's own sample sources
- Importation analysis at population level: "87% of
    imported cases came from Europe"

WHAT COUNTS AS "hospitalization":
POSITIVE: "cases admitted to our hospital", "patient
     was hospitalized", "required ICU admission"
NEGATIVE: "we enrolled patients from 4 hospitals" (
    study design, no hospitalization metadata)

CRITICAL DISTINCTION - PATIENT-LEVEL vs. POPULATION-
    LEVEL vs. STUDY DESIGN:
Patient metadata must describe individuals or a
    defined cohort in the current study. NOT
    patient metadata:
- Population-level case statistics: "5,093 cases
    reported in Amazonas"
- Epidemiological context: "the first confirmed case
     was reported in..."
- Study enrollment descriptions without actual
    patient values: "we enrolled patients from 4
    hospitals in Hong Kong"
- Study protocols: "participants who tested positive
     were placed in isolation"

CLASSIFICATION TRAPS - if a sentence matches a trap,
     it MUST be labeled NEGATIVE:

TRAP A (Previous/External Studies): Metadata from
    other studies, prior publications, external
    databases, or reference sequences.
```

```
Examples: "a postmortem pathological study found
    ...", "consistent with previously published
    evidence (Author, Year)"

TRAP B (Methodology/Variables): Variables or
    parameters mentioned without actual values for
    patients.
Examples: "we analyzed how severity was impacted by
    age" (no ages given), "the lack of association
    between G614 and hospitalization" (statistical
    finding about a variable, not actual patient
    data)
EXCEPTION: Figure/table captions that describe the
    study's own cohort data (e.g., "persistence of
    symptoms at 2 and 6 months after hospital
    admission for COVID-19", "Demographic and
    clinical characteristics of patients") ->
    POSITIVE.
EXCEPTION: Sentences describing sample collection
    that imply hospitalization (e.g., "collected
    from 126 cases admitted to our hospital") ->
    POSITIVE.

TRAP C (General Virology): Viral genetics, mutations
    , phylogenetics, or lab methods WITHOUT patient
    -level details.
CRITICAL: TRAP C should ONLY be triggered when the
    sentence is PURELY about the virus with NO
    reference to the study's own patients or sample
     origins. If the sentence mentions ANY of the
    following tied to the current study, do NOT
    trigger TRAP C - label POSITIVE instead:
- Geographic origin of the study's samples or
    sequences (country, city, region, hospital)
- A treatment given to the study's patients (e.g., "
    convalescent plasma therapy")
- Patient sample types or collection context that
    implies clinical metadata
If uncertain whether geography refers to the study's
     own samples vs. external references, consider
    it the study's own and label POSITIVE.

LOGICAL CONSTRAINT (MANDATORY):
If is_trap_triggered is true or trap_type is not "
    None", then patient_metadata_mentioned MUST be
    false. These fields must be consistent.

LABEL POSITIVE (TRUE):
The sentence reports or references patient metadata
    for the current study. This includes:
- Actual demographic values (ages, sex distribution)
- Clinical information (symptoms, severity, outcomes
    , treatments, hospitalization)
- Geographic origin of the study's samples or
    patients
- Lab results for the study's patients (test results
    , Ct values, viral load)
- Sample collection context that implies location or
     clinical setting
- Section/figure/table titles that name the study's
    geography or cohort characteristics

LABEL NEGATIVE (FALSE):
The sentence has no valid patient metadata for the
    study's own patients, or it falls into a trap
    above.

// USER PROMPT
Analyze the following PubMed extract. Apply your
    classification rules to determine if any
    patient metadata is mentioned, and extract the
    required information.
Extract: {raw_pubmed_text}
```

JSON Schema:

```
{
  "name": "covid_19_metadata_labeling",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "unique_id",
      "patient_metadata"
    ],
    "additionalProperties": false,
    "properties": {
      "unique_id": {
        "type": "string",
        "description": "The unique identifier as
            provided in the text"
      },
      "patient_metadata": {
        "type": "object",
        "required": [
          "patient_metadata_mentioned",
          "patient_metadata_citation"
        ],
        "additionalProperties": false,
        "properties": {
          "patient_metadata_mentioned": {
            "type": "boolean",
            "description": "Whether any patient
                metadata clearly associated with
                the study's sequenced cases is
                mentioned in the text"
          },
          "patient_metadata_citation": {
            "type": "string",
            "description": "The verbatim input text
                that supports the label for
                patient_metadata_mentioned. If
                patient_metadata_mentioned is false
                , this should be an empty string."
          }
        }
      }
    }
  }
}
```

A.8 Task 6: TNM Extraction Prompts and Schema

Prompt Structure:

```
// SYSTEM PROMPT
You are an expert oncological pathologist. Your task
    is to extract the TNM cancer stage from TCGA
    pathology reports (text provided)

You must analyze the provided clinical text and
    independently determine three specific
    classifications:
1. T (Primary Tumor Stage): Represents the size and
    extent of the main tumor.
2. N (Lymph Node Involvement): Represents the number
    and location of regional lymph nodes that have
    cancer.
3. M (Metastasis): Represents whether the cancer has
    spread to other parts of the body.

- If explicit staging is missing, use the
    macroscopic and microscopic descriptions (e.g.,
    tumor size in cm, number of positive lymph
    nodes) to infer the correct clinical stage
    based on standard oncological guidelines.

<Critical: Metastasis Detection>
DO NOT reflexively default to M0. Carefully search
    the ENTIRE report for ANY of the following M1
    indicators:
- Explicit "M1" or "pM1" notation anywhere in the
    text
- Tumor found in a distant organ (e.g., liver
    metastasis, bone metastasis, brain metastasis,
    contralateral lung)
- Phrases like "distant metastasis", "metastatic
    deposit in [organ]", "consistent with
    metastatic [cancer type]"
- Tumor in non-regional lymph node stations
- Peritoneal carcinomatosis, pleural metastasis,
    omental deposits
- Reports describing a biopsy of a SECONDARY site (
    not the primary tumor)
Only assign M0 if you have thoroughly checked and
    found NO evidence of distant spread.
</Critical>

// USER PROMPT
READ CAREFULLY: Pathology reports are often surgical
     and may explicitly stage the tumor (e.g., "pT2
    N1 Mx").
```

- If explicit staging is missing, use the macroscopic and microscopic descriptions (e.g., tumor size in cm, number of positive lymph nodes) to infer the correct clinical stage based on standard oncological guidelines.

JSON Schema:

```
{
  "name": "TNM_staging_reason_first",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "patient_filename",
      "primary_tumour_stage",
      "lymph_node_involvement",
      "metastasis"
    ],
    "additionalProperties": false,
    "properties": {
      "patient_filename": {
        "type": "string",
        "description": "The patient filename as
            provided in the text"
      },
      "primary_tumour_stage": {
        "type": "object",
        "required": [
          "evidence",
          "explicit_tnm_string",
          "reasoning",
          "T_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes from
                  the report relevant to T staging
                  (tumor size, depth of invasion,
                  organ involvement, explicit pT
                  notation)."
            },
            "description": "Extract ALL verbatim
                text from the report relevant to
                determining T stage BEFORE making a
                 classification."
          },
          "explicit_tnm_string": {
            "type": "string",
            "description": "If the report contains
                an explicit T stage (e.g. 'pT2', '
                T3', 'Stage IIIA'), copy it here
                verbatim. If none found, write '
                none'."
          },
          "reasoning": {
            "type": "string",
            "description": "Based on the evidence
                and any explicit staging, explain
                your reasoning for the T stage
                classification."
          },
          "T_stage": {
            "type": "string",
            "enum": [
              "T1",
              "T2",
              "T3",
              "T4"
            ],
            "description": "Your final T stage
                classification."
          }
        }
      },
      "lymph_node_involvement": {
        "type": "object",
        "required": [
          "evidence",
          "explicit_tnm_string",
          "lymph_node_counts",
          "reasoning",
          "N_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes from
                  the report about lymph nodes (
                  counts, stations, positivity,
                  explicit pN notation)."
            },
            "description": "Extract ALL verbatim
                text from the report relevant to
                determining N stage BEFORE making a
                 classification."
          },
          "explicit_tnm_string": {
            "type": "string",
            "description": "If the report contains
                an explicit N stage (e.g. 'pN1', '
                N2'), copy it here verbatim. If
                none found, write 'none'."
          },
          "lymph_node_counts": {
            "type": "string",
            "description": "Summarize the lymph node
                 counts: how many positive out of
                how many examined? e.g. '3/15
                positive'. Write 'not reported' if
                no counts given."
          },
          "reasoning": {
            "type": "string",
            "description": "Based on the evidence,
                explicit staging, and node counts,
                explain your reasoning for the N
                stage classification."
          },
          "N_stage": {
            "type": "string",
            "enum": [
              "N0",
              "N1",
              "N2",
              "N3"
            ],
            "description": "Your final N stage
                classification."
          }
        }
      },
      "metastasis": {
        "type": "object",
        "required": [
          "evidence_for_m1",
          "evidence_against_m1",
          "explicit_tnm_string",
          "reasoning",
          "M_stage"
        ],
        "additionalProperties": false,
        "properties": {
          "evidence_for_m1": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes
                  suggesting distant metastasis (
                  distant organ involvement, pM1,
                  non-regional nodes, secondary
                  site biopsies)."
            },
            "description": "Extract ALL verbatim
                text that could indicate M1. If
                nothing found, return empty array."
          },
          "evidence_against_m1": {
            "type": "array",
            "items": {
              "type": "string",
              "description": "Verbatim quotes
                  suggesting no distant metastasis
                  (negative distant workup, pM0,
                  confined to primary site)."
            },
            "description": "Extract ALL verbatim
```

```
                text that supports M0. If nothing
                found, return empty array."
            },
            "explicit_tnm_string": {
              "type": "string",
              "description": "If the report contains
                an explicit M stage (e.g. 'pM1', '
                M0', 'Mx'), copy it here verbatim.
                If none found, write 'none'."
            },
            "reasoning": {
              "type": "string",
              "description": "Weigh the evidence for
                and against M1. Explain your final
                classification."
            },
            "M_stage": {
              "type": "string",
              "enum": [
                "M0",
                "M1"
              ],
              "description": "Your final M stage
                classification."
            }
          }
        }
      }
    }
  }
}
```