

LLM Pros at #SMM4H-HeaRD 2025

Data Extraction Using Prompt Engineering And Structured Outputs (Task 1, 2, 3, 4, 5, 6)

Aatish Pradhan¹, Brian Habersberger¹, James Wade¹, Denver Dsouza¹, Nihal Paul¹,

¹Independent Researchers

aatishpr@buffalo.edu, habersberger@gmail.com, james.wade1221@gmail.com, denvereltondsouza@gmail.com, nihal91109@gmail.com

Abstract

Six tasks released for the SMM4H–HeaRD 2025 workshop were addressed with a unified large-language-model (LLM) pipeline that relies on prompt engineering, strictly enforced JSON schemas and lightweight rule sets. The pipeline utilizes no task-specific fine-tuning and can be practiced with minor modifications across a variety of data. The goal of this study was to demonstrate that general-purpose and widely available large language models (LLMs) are capable of understanding and extracting crucial health information. The systems achieved the highest-ranked submission scores on the official leaderboard for Task 2 (non-medical substance use), all three Subtasks of Task 4 (insomnia), and Subtask 2 of Task 5 (foodborne outbreak entity extraction). A detailed workflow on insomnia (Task 4) illustrates how sleep-difficulty rules, daytime-impairment rules and medication lists interacted. Shorter descriptions are provided for the remaining five tasks. On the test data sets, the systems obtained an F1-score of 0.4234 for Task 2, an F1-score of 0.9670 for Task 4 (Subtask 1), an F1-score of 0.9064 for Task 4 (Subtask 2a), an F1-score of 0.6822 for Task 4 (Subtask 2b), and an average score of 0.576 for Task 5 (Subtask 2).

Introduction

Social-media mining for health (SMM4H) tasks typically demand nuanced understanding of domain language, lengthy context windows and precise evidence localization. Conventional approaches employ domain-specific BERT variants or sequence-to-sequence fine-tuning on each task. Such methods incur substantial GPU cost and require hundreds to thousands of supervised examples. It is important to note that while previous attempts have been made to tune results using GPT-4 (Mukans and Barzdins 2024), recent advances in large-language models (LLMs) enable constrained decoding, in which the model is compelled to emit syntactically valid JSON that follows a user-supplied schema. When key decision logic is embedded in this schema, the need for explicit fine-tuning can be avoided.

The present study evaluates this idea on six heterogeneous tasks released for SMM4H–HeaRD 2025: adverse drug events (ADE; multilingual), impacts of non-medical substance use, dementia caregiver identification, insomnia

detection in clinical notes, food recall / outbreak monitoring, and vaccine adverse events. **The main contributions are:**

1. **A unified framework that scored among the top three of the six tasks without any task-specific fine-tuning.**
2. **A detailed analysis of insomnia classification demonstrating how domain knowledge (drug lists) captured in the prompt transfers into improved evidence extraction.**
3. **An analysis that characterizes discrepancies between labels and predictions in terms of entity-relationship specification requirements for LLM-based extraction tasks.**

Recent work in social media mining for health has leveraged both traditional transformer-based models and large language models (LLMs) to address tasks such as adverse drug event (ADE) extraction, classification of mental health symptoms, and identification of medical conditions in children. Teams have adopted diverse strategies including encoder-decoder architectures (e.g., BART, T5), RoBERTa ensembles, and fine-tuned domain-specific models. Data augmentation—through paraphrasing, random shuffling, or synthetic generation using LLMs like ChatGPT—has been widely used to enhance performance on imbalanced and noisy datasets. Some approaches incorporated cross-task training and knowledge distillation to transfer domain knowledge across related problems, while others demonstrated the effectiveness of lightweight or few-shot learning in low-resource settings. Collectively, these methods underscore the growing utility of LLMs and robust pretraining strategies in extracting actionable health insights from social media content (Klein et al. 2023; Ta et al. 2024; Fan, Yang, and Cao 2024; Wasi and Rahman 2024).

While some loosely-followed string-encoded structure was possible to articulate and evoke from LLMs previously, strict structured output capabilities were introduced by OpenAI in August 2024 in the model *gpt-4o-2024-08-06* (OpenAI 2022). Structured outputs are useful in tasks that involve extracting multiple types of entities and their relationships from unstructured text. To use schema-constrained generation effectively, both the prompt design and the associated JSON schema must comprehensively define the relevant domain-specific information. The schema should

clearly specify the entity types, their relationships, and their allocation framework, providing explicit guidance to the model on how information should be organized. The order in which entities and relationships are defined within the schema also contributes to the model’s understanding of the ontology, by reflecting the hierarchical or sequential importance of information elements within the task. At the same time, the schema must remain flexible enough to allow the language model to produce outputs that are both syntactically correct and contextually accurate. Incorporating an entity-relationship framework further supports the model by explicitly mapping the semantic connections among entities. This framework assists the model in interpreting and aligning its outputs with the underlying ontology of the task, ensuring consistency and relevance in domain-specific information extraction.

Prompt engineering (here, inclusive of schema design) allows us to guide LLMs to generate structured outputs without requiring large-scale labeled datasets. By carefully designing prompts that specify the desired information format, we can directly extract relevant entities and classifications from unstructured text. This approach takes advantage of the pre-trained knowledge stored in LLMs, which have been trained on vast amounts of diverse textual data, enabling them to recognize patterns and infer relationships even with limited domain-specific supervision.

For each task, we combined close reading of the task descriptions with systematic examination of the provided training labels to develop our structured output schemas. Since human-annotated labels (ground truth labels) may contain inconsistencies and explicit criteria for defining entity relationships were not provided, we inferred the implicit entity-relationship frameworks that appeared to underlie the annotators’ labeling decisions. These frameworks guided our schema design to align with the patterns observed in the training data.

In summary, our methodology has been applied to all six tasks within this workshop, leveraging the structured output framework to mitigate the need for training and fine tuning models to obtain accurate results. Ultimately, with this strategy we aimed to achieve robust and accurate outputs, thereby advancing the state-of-the-art in social media mining for health-related research within the context of the #SMM4H workshop (Klein et al. 2025).

Methodology

By defining domain-specific terms (e.g., In task 4 - difficulty sleeping and daytime impairment) and listing medication groups, the prompt guides the model to generate structured outputs that are subsequently processed using rule-based label assignment. A structured schema that explicitly allocates each diagnostic criterion into two clinically meaningful entities—Difficulty Sleeping and Daytime Impairment—while rigorously linking each extracted criterion to its supporting citation in the text. This clear separation and explicit mapping of entities to evidence ensures a consistent interpretative framework, minimizing ambiguity of the unstructured text and improving reproducibility in model outputs.

By formalizing both the entities and their relationships to source evidence through this schema, the LLM can reliably disentangle complex clinical narratives and deliver structured, machine-readable extractions aligned with clinical guidelines. This entity relationship-driven framework played a central role in driving the method’s robust performance without the need for task-specific fine-tuning. Task 2 required low reasoning and was addressed using GPT-4o. Tasks with complex entity relationships used GPT o3-mini.

Architecture

- **Raw text** from the social media provided is used
- **Prompt assembly:** Three segments are concatenated:
 - (SYSTEM) high-level domain instruction;
 - (USER) the cleaned document;
 - (SCHEMA)
- **Model call** - All tasks (except task 2) use GPT-o3-mini. Task 2 uses GPT-4o
- **Validation & rule based labeling** (if-else statements were applied to the structured LLM responses)

The architecture was applied consistently across Tasks 1, 3, 4, 5, and 6. Task 2, while similar in its overarching approach, incorporated an additional refinement through the integration of the *Pydantic* library (Pydantic Developers 2025). This enabled token-level labeling based on precise string matching, which was particularly effective in addressing the challenges associated with identifying brief impact phrases within Reddit posts related to nonmedical substance use.

Task 1: Detection of adverse drug events in multilingual and multi-platform social media posts

Task 1 centers on extracting descriptions of adverse drug effects from social media posts. The aim is to detect and classify instances where users explicitly report negative outcomes and to establish a direct connection between these adverse effects and any cited causes, ensuring that only clearly attributable experiences are captured. A representative excerpt of the prompt used for Task 1 is provided in Listing 1.

Task 2: Extraction of clinical and social impacts of non-medical substance use from Reddit

The pipeline leveraged OpenAI’s GPT-4o language model to identify short, high-precision impact phrases from raw post content. Posts were pre-annotated in a token-level TSV format, which was parsed to reconstruct the original post texts alongside their token metadata. For impact extraction, each unique post was submitted to the language model using an instruction-tuned prompt that emphasized the identification of the shortest possible phrases denoting either clinical or social consequences. A structured schema, defined using the *Pydantic* library, was employed to enforce consistent formatting of model responses and to categorize each identified phrase under “Clinical Impacts,” “Social Impacts,” or “Other.” Each predicted phrase was subsequently aligned

Listing 1: Condensed system prompt for Task 1. Full details are provided in Appendix Section

```
1 SYSTEM: You are an assistant
    specializing in identifying
    descriptions of adverse drug effects
    in social media posts. Note that the
    text may be in any of English, French
    , Russian, or German. There should be
    a clear connection drawn between the
    negative effect and any potential
    causes, not just a report of both a
    negative effect and a potential cause
    . For example, in the text "My
    stomach hurts whenever I overeat, but
    I'm eating less often because of
    that new pill I'm taking". In this
    example, there is no clear connection
    between "that new pill" and the
    stomach pain, although it may be
    considered ambiguous.
2 USER: <<< INSERT SOCIAL MEDIA MESSAGE
    HERE >>>
3 RESPONSE_FORMAT: task1_schema
```

with the corresponding tokenized post, and token-level labels were assigned based on string matches between model-extracted spans and original tokens.

System Prompt: *“You are an expert in extracting *specific phrases* from Reddit posts that indicate clinical or social impacts of nonmedical substance use. Only return the *shortest possible phrase* that represents the impact. Do not include surrounding context or full sentences. If no impact is present, return an empty list. Instances in the clinical impacts category describe the clinical effects, consequences, or impacts of substance use on individuals’ health, physical condition, or mental well-being. Instances the social impacts describe the societal, interpersonal, or community-level effects, consequences, or impacts of nonmedical substance use. These impacts may include social relationships, community dynamics, or broader social issues. For example, ‘Truth is, I was in the psych ward from Tuesday afternoon until this evening.’ Should identify ‘psych ward’ for Clinical Impact.”*

Algorithm 1 highlights the workflow using the Pydantic library and OpenAI

Task 3: Detection of dementia family caregivers on Twitter

This binary classification task aims to automatically identify English-language tweets reporting a family member with dementia versus general mentions of dementia.

A representative excerpt of the prompt used for Task 3 is provided in Listing 2.

Task 4: Detection of insomnia in clinical notes

Task 4 focuses on the extraction of insomnia-related diagnostic criteria from anonymized clinical notes. The objective is to identify and classify evidence of sleep difficulty

Algorithm 1: Extract and Label Impact Phrases Using Pydantic and OpenAI for Task 2

Input: Annotated TSV file of Reddit posts
Output: TSV file with tokens labeled as clinical or social impacts

- 1: Define Pydantic schema:
 - ImpactType: Enum with Clinical, Social, and Other
 - ImpactSpan: Includes an excerpt and its impact type
 - ImpactExtraction: A list of ImpactSpan objects
 - 2: Parse the annotated TSV file into a token-level DataFrame
 - 3: Extract unique Reddit post texts
 - 4: **for** each post text **do**
 - 5: Use OpenAI API to extract impact excerpts using the ImpactExtraction schema
 - 6: Match and label tokens in the post if they correspond to extracted impact phrases
 - 7: **end for**
 - 8: Combine all labeled post DataFrames
 - 9: Export labeled tokens to a TSV file
-

Listing 2: Condensed system prompt for Task 3. Full details are provided in Appendix Section

```
1 SYSTEM: You are an assistant
    specializing in identifying
    descriptions
2 of dementia and other illnesses in
    social media posts from sources such
    as Twitter.
3 For this task, label all causes of
    dementia (e.g., Alzheimer’s disease)
    as "Dementia".
4 Exclude non-dementia related illnesses (
    for example, Parkinson’s disease).
5 USER: <<< INSERT TWEET HERE >>>
6 RESPONSE_FORMAT: task3_schema
```

and daytime impairment, and, where applicable, to associate these findings with related medication mentions.

A representative excerpt of the prompt used for Task 4 is provided in Listing 3.

Task 5: Detection and extraction of food recalls and foodborne disease outbreaks in online news articles

The objective is to both classify relevant sentences—differentiating between Food Recall, Foodborne Disease Outbreak, or Neither—and to extract detailed event information. In particular, systems are expected to identify key entities such as the target organization, detailed product information, the disease or infection name, specifics of the safety incident, and an indication of the number of people affected. A representative excerpt of the prompt used for Task 5 is provided in Listing 4.

Listing 3: Condensed Task 4 system prompt. Full details are provided in Appendix Section

```

1 SYSTEM: You are a clinical NLP assistant
  specialized in identifying insomnia-
  related diagnostic criteria from
  anonymized clinical notes.
2 Definitions: DifficultySleeping := {
  initiate sleep | maintain sleep |
  early awakening | "insomnia"}
3 DaytimeImpairment := {fatigue | impaired
  concentration | sleep
  dissatisfaction | daytime sleepiness
  | mood disturbance}
4 USER: <<< INSERT CLINICAL NOTE HERE >>>
5 RESPONSE_FORMAT: task4_schema

```

Listing 4: Condensed Task 5 system prompt. Full details are provided in Appendix Section

```

1 SYSTEM: You are a text analysis
  assistant specializing in identifying
  information related to food product
  health concerns. Analyze the text,
  providing a text_label as a
  categorization. Additionally,
  identify any organizations
  responsible for the issue (typically
  food manufacturers), products, causes
  (e.g. "contamination with Salmonella
  ", "failure to follow industry
  hygiene standard", etc.), diseases (e
  .g. "Salmonella infection"), number
  of people affected (report as a
  string, may be approximate or a range
  ), and locations (e.g. cities, states
  , regions, etc.).
2 USER: <<< INSERT TEXT HERE >>>
3 RESPONSE_FORMAT: task5_schema

```

Task 6: Detection of adverse vaccine events on Reddit

The objective is to ensure that only posts containing first-hand accounts of adverse reactions to herpes zoster vaccines—where the narrator explicitly describes personal symptoms and experiences—are identified, while general discussions or speculative commentary about vaccine safety are excluded. A representative excerpt of the prompt used for Task 6 is provided in Listing 5

This structured approach eliminates the need for task-specific fine-tuning, relying instead on the model’s pre-trained capabilities and the rigor of enforced JSON schemas to ensure consistency and precision in identifying key information from text.

Results

Tables 1 and 2 summarize the performance of our system across all six tasks and provide a benchmark comparison against competing systems, respectively. For Task 5 Subtask 2, scores reflect a weighted word-overlap similarity across six entity fields; an asterisk (*) indicates this is not a stan-

Listing 5: Condensed system prompt for Task 6. Full details are provided in Appendix Section

```

1 SYSTEM: You are an assistant
  specializing in identifying
  descriptions
2 of adverse vaccine reactions in social
  media posts.
3
4 <Some helpful information in identifying
  certain vaccines> A zoster vaccine
  reduces the incidence of herpes
  zoster (shingles) caused by
  reactivation of the varicella zoster
  virus (also responsible for
  chickenpox). Shingles provokes a
  painful rash with blisters and may
  lead to chronic pain (postherpetic
  neuralgia) and additional
  complications. Older or
  immunosuppressed individuals are
  especially vulnerable. Two zoster
  vaccines (Shingrix and Zostavax) have
  been approved for people over 50;
  note that Zostavax was discontinued
  in the United States in November
  2020. Additionally, a varicella
  vaccine is used to prevent diseases
  caused by the same virus. </Some
  helpful information in identifying
  certain vaccines>
5 USER: <<< INSERT TEXT HERE >>>
6 RESPONSE_FORMAT: task6_schema

```

dard F1-score. Table 1 presents precision, recall, and F1-score metrics achieved by our approach for each task and subtask, along with notes highlighting cases where our system achieved the best-in-class performance or notable rankings. In contrast, Table 2 reports the median and mean performance scores from all participating systems for each task and subtask, offering context to evaluate the relative effectiveness of our method. Together, these tables provide a comprehensive overview of system-level performance and demonstrate the competitiveness and robustness of our structured output framework across diverse information extraction and classification tasks. Statistical analysis is omitted, as the focus is on deterministic, schema-constrained LLM outputs rather than stochastic variability.

| Task | Subtask | Precision | Recall | F1-score | Notes |
|------|-----------|-----------|--------|----------|---------------|
| 1 | — | 0.4683 | 0.8029 | 0.5915 | — |
| 2 | — | — | — | 0.4234 | Best-in-class |
| 3 | — | 0.8020 | 0.1160 | 0.2030 | — |
| 4 | Subtask1 | — | — | 0.9670 | Best-in-class |
| | Subtask2A | — | — | 0.9064 | Best-in-class |
| | Subtask2B | — | — | 0.6822 | Best-in-class |
| 5 | Subtask1 | — | — | 0.8080 | Ranked 3rd |
| | Subtask2 | — | — | 0.5760* | Best-in-class |
| 6 | — | 0.8430 | 0.9240 | 0.8820 | — |

Table 1: Performance of our system across tasks

| Task Subtask | Metric | Precision | Recall | F1-score |
|--------------|--------|-----------|--------|----------|
| 1 | Median | 0.6172 | 0.6312 | 0.6268 |
| | Mean | 0.5440 | 0.5661 | 0.5394 |
| 3 | Median | 0.9460 | 0.9690 | 0.9530 |
| | Mean | 0.9250 | 0.8920 | 0.8850 |
| 4-Subtask1 | Median | 0.8687 | 0.8400 | 0.9348 |
| | Mean | 0.8771 | 0.8526 | 0.9130 |
| 4-Subtask2A | Median | 0.6922 | 0.6500 | 0.8182 |
| | Mean | 0.7170 | 0.6726 | 0.7885 |
| 4-Subtask2B | Median | 0.4459 | 0.5140 | 0.4873 |
| | Mean | 0.3858 | 0.4231 | 0.4430 |
| 5-Subtask1 | Median | 0.8670 | 0.8240 | 0.8240 |
| | Mean | 0.8870 | 0.8548 | 0.8535 |
| 5-Subtask2 | Median | – | – | 0.3260* |
| | Mean | – | – | 0.3368* |
| 6 | Median | 0.9220 | 0.9720 | 0.9440 |
| | Mean | 0.9160 | 0.9610 | 0.9380 |

Table 2: Competitor benchmark (median and mean) scores across tasks

Structured LLM Outputs as a Lens for Task Definition Using Entity Relationship Frameworks

In this work, high-performance prompts were coupled with a strictly enforced structured response format to improve extraction accuracy and ensure consistency in information retrieval from large language models.

This approach reveals subtle variations in the interpretation of social media texts. While human interpretations can vary across different individuals and contexts, our system ensure that the LLM consistently follows a uniform chain of reasoning. While examples provided here are indicative of broader trends, it is important to acknowledge that they may not capture the full scope or frequency of these phenomena in the dataset. A comprehensive assessment of differences in interpretation is beyond the scope of this manuscript.

A closer look at the structured outputs, for each task, brings to light some of these nuances that may be attributable to differing human interpretation of these texts. Even in superficially similar cases, the structured responses highlighted subtle differences in annotation practices and interpretative standards. Through these examples, we showcase that when provided with clearly defined entity relationship frameworks, the LLMs consistently generate outputs that are useful for the intended analytical purposes. Subsequent sections present detailed descriptions and examples to further clarify these observations and improving the effective deployment of LLMs in research contexts.

Task 1 validation data set example: ID en_346, provides a compelling example of the challenges inherent in annotating adverse drug event mentions when task definitions remain ambiguous. This was originally assigned a label of 0 (no mention of Adverse Drug Event) and the LLM predicted an output of 1. In this example, the text – “@USER_____ I an had incident with an imodium tablet. It’s safe to say I will never be taking one of them again. Hope you are okay.

xx” – clearly references negative outcomes associated with a specific drug.

The LLM’s output, structured as, { “negative_effects_and_causes”: [{ “adverse_effect”: “Had an incident after taking an imodium tablet”, “person_experiencing_effect”: “author of the text”, “attributed_causes”: [{ “name_of_cause”: “drug, pharmaceutical, pill, etc.”, “degree_of_attribution”: “clear and direct” }] }] } – exemplifies both the potential and example of where a criterion is not defined of structured output approaches. The complexity of this output demonstrates that, without clearly defined boundaries for what constitutes an “adverse drug event mention”, it is difficult to determine how the text should be labeled. Establishing more explicit criteria—such as specifying whether the mention should correspond to a particular symptom or condition—is necessary to ensure that the LLM’s outputs align consistently with the intended task.

Task 3 validation data set example: ID 1393792840312971265, illustrates the nuanced challenges associated with classifying complex relational dynamics in health-related social media texts. In this example, the text – “My husband’s grandmother has dementia and accused me of assault – can I skip next family gathering? <https://t.co/qq4vKfpbst>” – was originally annotated with a ground truth label of 0, yet the large language model predicted a label of 1. The LLM’s output, structured as, { “relationships_to_illness”: [{ “identity_of_ill_person”: “husband’s grandmother”, “type_of_illness”: “Dementia”, “relationship_to_text_author”: “Family member (explicitly stated)” }] }, exposes a fundamental discrepancy in the interpretation of “family member” criteria. This divergence stems from the inherent variability in how familial relationships are defined across different annotators and/or cultural contexts, thereby complicating the creation of a universally applicable labeling schema. These examples underscore the broader challenge of task formulation, where ambiguities in defining relational concepts can introduce substantial labeling inconsistencies, whether for subject matter experts or LLMs.

Task 5 validation data set example: The examples encountered in this task illustrate several key considerations common in entity recognition, allocation, and relationship extraction frameworks. In instances where a text referenced multiple event types—such as a foodborne disease outbreak and a food recall—the annotation process involved associating entities with the appropriate event class. Observations suggest that a consistent mapping to a primary event type, such as “food recall”, can help align outputs with the task objectives. Similarly, when a single incident included references to multiple time points, entity anchoring decisions—such as selecting either the initial occurrence or the most recent mention—play an important role in ensuring consistent temporal representations, depending on the intended analytical focus.

The extraction of information for fields such as “location” encompassed varying geographic granularities, including countries, states, cities, and regions, as well as different types of location references, such as where food was pur-

chased, where illness occurred, or where the company owner was arrested. Ambiguity in field population arose from diverse location mentions and the absence of clear criteria, resulting in inconsistent extraction.

Docid 2882, illustrates a notable discrepancy between the ground truth annotation and the LLM’s prediction, thereby highlighting the challenges in accurately capturing the nuances of outbreak-related narratives. The text – *The table below shows outbreak investigations being managed by FDA’s CORE Response Teams. The investigations are in a variety of stages. Some outbreaks have limited information with investigations ongoing, others may be near completion. The Food and Drug Administration will issue public health advisories for outbreak investigations that have resulted in specific, actionable steps for consumers – such as throwing out or avoiding specific foods – to take to protect themselves, according to the outbreak table page. Not all recalls and alerts result in an outbreak of foodborne illness. Outbreak investigations that do not result in specific, actionable steps for consumers may or may not conclusively identify a source or reveal any contributing factors, according to CORE’s outbreak table page. If a sources and/or contributing factors are identified that could inform future prevention, FDA commits to providing a summary of those findings, according to CORE officials. Click here to visit the FDA page that has links to outbreak information.* While the text has no specific description of any indicents but instead makes reference to an undisclosed table and general FDA practice, the ground truth label was assigned as “Foodborne Disease Outbreak”, whereas the LLM correctly categorized the text as “Neither.”

The LLM’s structured output as, { ‘events’: [{ ‘incident_timeline’: ‘Present’, ‘text_label’: ‘Neither’, ‘extracted_entities’: { ‘organizations’: [{ ‘name’: ‘Food and Drug Administration’, ‘roles’: [‘Other’]}, { ‘name’: “FDA’s CORE Response Teams”, ‘roles’: [‘Other’]}], ‘product’: [], ‘cause’: [], ‘disease’: [], ‘number_of_people_affected’: [], ‘origin_location’: [], ‘affected_location’: [] }] }

The LLM’s structured output further reinforces this interpretation by clearly outlining the absence of specific adverse events: it identifies entities such as the Food and Drug Administration and FDA’s CORE Response Teams under the organizational category while leaving key event-related fields (e.g., product, cause, disease, number_of_people_affected, origin_location, affected_location) empty. This case exemplifies situations in which the LLM’s output appears to better align with the goal of the task.

Task 6 validation data set example: ID 2833 provides an example of the challenges involved in systems that capture the adverse reactions to shingles vaccinations, as well as the potential for automated systems to surpass manual annotations in capturing salient details. In this case, the text references a pneumonia vaccine – *Had my pneumonia vaccine done today and let me tell you, the anticipation was way worse than the actual shot. It was over in a second and I didn’t even flinch. Shoutout to all the healthcare workers who make getting vaccinated a breeze!* – which describes a pneumonia vaccine-related event, with the ground truth label being assigned as “Yes”. Conversely, the LLM predicted a “No” label, supported by a structured output that

meticulously delineates no adverse reaction. Specifically, the LLM’s structured output, { ‘social_media_post_text’: [{ ‘text_type’: ‘Social media post content (Likely original post)’, ‘text’: ‘Had my pneumonia vaccine done today and let me tell you, the anticipation was way worse than the actual shot. It was over in a second and I didn’t even flinch. Shoutout to all the healthcare workers who make getting vaccinated a breeze!’}, ‘adverse_reaction_to_vaccine’: None] }] encapsulates the adverse reaction to vaccine as none.

Ultimately, establishing clear entity relationship frameworks not only guides large language models in generating structured outputs but also enhances consistency and clarity for human annotators when assigning labels, thereby improving reliability across both automated and manual annotation processes. These observations do not indicate oversights but instead underscore the inherent complexity of labeling social media or other unstructured data, as well as the range of criteria that must be carefully considered when applying LLMs to these tasks.

Conclusion

This work investigated how modern LLMs can extract vital health information from the noisy and unstructured texts found in social media. We introduced a unified framework that uses structured output schemas—specifically predefined JSON formats—combined with optimized prompt strategies to generate precise, consistent, and machine-readable outputs.

We applied this approach across diverse tasks, including adverse drug event detection, insomnia identification, food recall extraction, and vaccine adverse event monitoring. This approach obviates the need for task-specific fine-tuning by leveraging the model’s pre-trained knowledge. Our framework demonstrated competitive performance by guiding large language models through domain-adapted prompts and schema constraints.

A central insight of this study is that clearly defined entity-relationship frameworks are key to ensuring consistency in both model-generated outputs and human annotations. By embedding these frameworks into schema-constrained prompts, LLMs maintain a stable chain of reasoning and produce reproducible outputs across diverse inputs. This enables LLMs to effectively automate much of the data labeling process, allowing human annotators to focus primarily on reviewing ambiguous or edge cases—streamlining information extraction and improving overall efficiency.

Overall, our findings highlight the value of structured output specifications—not only in format but also in guiding entity and relationship assignment—for improving accuracy, reproducibility, and consistency in health-related information extraction from informal text. The proposed approach is broadly applicable to other domains requiring structured extraction from unstructured data.

Acknowledgments

This study was self-funded by the team members. The authors have no conflicts of interest to disclose.

References

- Fan, Y.; Yang, D.; and Cao, L. 2024. CTYUN-AI@SMM4H-2024: Knowledge Extension Makes Expert Models. 5–9.
- Klein, A. Z.; Banda, J. M.; Guo, Y.; Schmidt, A. L.; Xu, D.; Amaro, J. I. F.; Rodriguez-Esteban, R.; Sarker, A.; and Gonzalez-Hernandez, G. 2023. Overview of the 8th Social Media Mining for Health Applications (#SMM4H) Shared Tasks at the AMIA 2023 Annual Symposium. medRxiv [Preprint]. Update in: J Am Med Inform Assoc. 2024 Apr 3;31(4):991-996. doi: 10.1093/jamia/ocae010. PMID: 37986776; PMCID: PMC10659479, medRxiv:2023.11.06.23298168.
- Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Gryboski, L.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Mazzotti, D.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.
- Mukans, E.; and Barzdins, G. 2024. RIGA at SMM4H-2024 Task 1: Enhancing ADE discovery with GPT-4. 23–27.
- OpenAI. 2022. Structured Outputs. *openai.com*.
- Pydantic Developers. 2025. Documentation for version: v2.11.3. <https://docs.pydantic.dev/latest/>. Accessed: 2025-04-07.
- Ta, T.; Rahman, A. B. S.; Najjar, L.; and Gelbukh, A. 2024. ThangDLU at #SMM4H 2024: Encoder-decoder models for classifying text data on social disorders in children and adolescents. 1–4.
- Wasi, A. T.; and Rahman, S. 2024. DILAB at #SMM4H 2024: RoBERTa Ensemble for Identifying Children’s Medical Disorders in English Tweets.

Appendix

Task 1 schema

```
{“name”:“adverse_effect_detection”, “strict”:true,
“schema”:{ “type”: “object”, “properties”:{ “negative_effects_and_causes”:{ “type”: “array”, “description”:
“Negative effects and causes in the text. Only include any directly reported experiences, not broad advisories or speculation about what a drug or other cause may be associated with.”, “items”:{ “type”: “object”, “properties”:{ “adverse_effect”:{ “type”: “string”, “description”: “A description of any adverse or negative experience, symptom, etc. in the text”}, “person_experiencing_effect”:{ “type”: “string”, “enum”:[ “author of the text”, “person familiar to author of the text (friend, family, coworker, etc.)”, “no specific person”, “other”]}, “attributed_causes”:{ “type”: “array”, “description”: “A list of any causes associated with the adverse effect”, “items”:{ “type”: “object”, “required”:[ “name_of_cause”, “degree_of_attribution”], “additionalProperties”:false, “properties”:{ “name_of_cause”:{ “type”:
```

```
“string”, “enum”:[ “illness”, “person”, “drug, pharmaceutical, pill, etc.”, “other”]}, “degree_of_attribution”:{ “type”: “string”, “enum”:[ “clear and direct”, “strong but indirect”, “ambiguous”]}]}}, “required”:[ “adverse_effect”, “attributed_causes”, “person_experiencing_effect”], “additionalProperties”:false}}}, “required”:[ “negative_effects_and_causes”], “additionalProperties”:false}}
```

Task 3 schema

```
{ “name”: “illness_relationship”, “strict”:true, “schema”:{ “type”: “object”, “properties”:{ “relationships_to_illness”:{ “type”: “array”, “description”: “A list of relationships to individuals with illnesses.”, “items”:{ “type”: “object”, “properties”:{ “identity_of_ill_person”:{ “type”: [ “string”, “null”], “description”: “The name (or other identity) of the person who has the illness.”}, “type_of_illness”:{ “type”: “string”, “description”: “The type of illness the person has.”, “enum”:[ “Dementia”, “Illness other than Dementia”]}, “relationship_to_text_author”:{ “type”: “string”, “description”: “The relationship of the ill person to the author of the text.”, “enum”:[ “Parent”, “Sibling”, “Grandparent”, “Aunt or Uncle”, “Step-parent”, “Step-sibling”, “Brother or sister-in-law”, “Other family member”, “Non-family member”, “Unclear or no relationship described”]}}, “required”:[ “identity_of_ill_person”, “type_of_illness”, “relationship_to_text_author”], “additionalProperties”:false}}}, “required”:[ “relationships_to_illness”], “additionalProperties”:false}}
```

Task 4 schema

```
{“name”：“clinical_notes_criteria”, “strict”:true,
“schema”:{ “type”：“object”, “properties”:{
“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”]}, “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”}}}}, “required”:[“difficulty_sleeping_identified_criteria”, “day-time_impairment_identified_criteria”], “additionalProperties”: false}}

Task 5 schema

```
{
  "name": "analyze_food_safety_text",
  "strict": true,
  "schema": {
    "type": "object",
    "required": [
      "events"
    ],
    "additionalProperties": false,
    "properties": {
      "events": {
        "type": "array",
        "description": "List of events",
        "items": {
          "type": "object",
          "required": [
            "incident_timeline",
            "text_label"
          ],
          "extracted_entities": false,
          "properties": {
            "incident_timeline": {
              "type": "string",
              "description": "Some presently occurring incidents may be compared to incidents in the past, use this field to note where on the timeline a given incident occurred.",
              "enum": [
                "Present",
                "Past"
              ],
              "text_label": {
                "type": "string",
                "enum": [
                  "Food Recall",
                  "Foodborne Disease Outbreak",
                  "Food Recall and Foodborne Disease Outbreak",
                  "Neither"
                ]
              },
              "extracted_entities": {
                "type": "object",
                "required": [
                  "organizations",
                  "product",
                  "cause",
                  "disease",
                  "number_of_people_affected",
                  "origin_location",
                  "affected_location"
                ],
                "additionalProperties": false,
                "properties": {
                  "organizations": {
                    "type": "array",
                    "description": "List of organizations liable or responsible for the cause of the food safety incident.",
                    "items": {
                      "type": "object",
                      "required": [
                        "name",
                        "roles"
                      ],
                      "additionalProperties": false,
                      "properties": {
                        "name": {
                          "type": "string",
                          "description": "Name of the organization."
                        },
                        "roles": {
                          "type": "array",
                          "description": "List of roles or responsibilities of the organization.",
                          "items": {
                            "type": "string",
                            "enum": [
                              "Origin of foodborne disease outbreak",
                              "Issuer of food recall",
                              "Other"
                            ]
                          }
                        },
                        "product": {
                          "type": "array",
                          "description": "List of food products involved.",
                          "items": {
                            "type": "string"
                          }
                        },
                        "cause": {
                          "type": "array",
                          "description": "List of food safety incident causes, usually contamination.",
                          "items": {
                            "type": "string"
                          }
                        },
                        "disease": {
                          "type": "array",
                          "description": "List of diseases caused by food safety incident.",
                          "items": {
                            "type": "string"
                          }
                        },
                        "number_of_people_affected": {
                          "type": "array",
                          "description": "List of numbers of people affected by food safety incidents. This may be more complex than simply a number, such as 'at least 200 sick with 5 confirmed deaths'",
                          "items": {
                            "type": "string"
                          }
                        },
                        "origin_location": {
                          "type": "array",
                          "description": "List of locations of the origin of the food safety incident.",
                          "items": {
                            "type": "string"
                          }
                        },
                        "affected_location": {
                          "type": "array",
                          "description": "List of locations of any areas affected by the food safety incident.",
                          "items": {
                            "type": "string"
                          }
                        }
                      }
                    }
                  }
                }
              }
            }
          }
        }
      }
    }
  }
}
```

Task 6 schema:

```
{
  "name": "adverse_reactions_schema",
  "strict": true,
  "schema": {
    "type": "object",
    "properties": {
      "social_media_post_text": {
        "type": "array",
        "description": "If the post appears to be in the form 'title': 'content',
```

```
identify these sub-segments of the post when analyzing.",
        "items": {
          "type": "object",
          "properties": {
            "text_type": {
              "type": "string",
              "description": "The type of text being analyzed",
              "enum": [
                "Social media post title",
                "Social media post content (Likely original post)",
                "Social media post content (Likely reply to original post)",
                "Other or unclear"
              ]
            },
            "text": {
              "type": "string",
              "description": "Repeat verbatim this portion of the social media post",
              "adverse_reaction_to_vaccine": {
                "type": [
                  "array",
                  "null"
                ],
                "description": "Adverse reactions described in the social media posting",
                "items": {
                  "type": "object",
                  "properties": {
                    "symptoms": {
                      "type": "string",
                      "description": "The symptoms experienced by the person affected."
                    },
                    "person_affected": {
                      "type": "string",
                      "description": "The person affected by the adverse reaction.",
                      "enum": [
                        "Original poster",
                        "Person replying to original post",
                        "Family member of post author",
                        "Other"
                      ]
                    },
                    "associated_cause": {
                      "type": "object",
                      "properties": {
                        "type_of_cause": {
                          "type": "string",
                          "enum": [
                            "Specified vaccine",
                            "Unspecified vaccine",
                            "Virus",
                            "Bacterium",
                            "other"
                          ]
                        },
                        "name_of_specified_vaccine": {
                          "type": "string",
                          "enum": [
                            "null"
                          ],
                          "description": "If type_of_cause is 'Specified vaccine', describe the vaccine here. Otherwise use null."
                        },
                        "required": [
                          "type_of_cause",
                          "name_of_specified_vaccine"
                        ],
                        "additionalProperties": false
                      }
                    },
                    "required": [
                      "person_affected",
                      "associated_cause",
                      "symptoms"
                    ],
                    "additionalProperties": false
                  }
                },
                "required": [
                  "text_type",
                  "text",
                  "adverse_reaction_to_vaccine"
                ],
                "additionalProperties": false
              }
            }
          }
        },
        "required": [
          "social_media_post_text"
        ],
        "additionalProperties": false
      }
    }
  }
}
```