Team Gazoo! at #SMM4H-HeaRD 2026: Zero-Training NER via Iterative LLM Prompt Self-Optimization for Opioid Impact Span Detection

Diego Estuar
Cedars-Sinai Medical Center
Los Angeles, CA, USA
Claremont Graduate University
Claremont, CA, USA
diego.estuar@cshs.org
diego.estuar@cgu.edu

Abstract

This paper describes the system submitted by Team Gazoo! for Task 7 of the #SMM4H-HeaRD 2026 shared task on detecting self-reported clinical and social impacts of non-medical opioid use in social media text. We present a zero-training, prompt-only approach that uses a large language model (GPT-5.4) with structured few-shot prompting and autonomous, iterative rule optimization. Our system encodes a domain-specific entity ontology, three core decision rules, and 65 cognitively organized few-shot examples into a single prompt, with BIO constraint enforcement applied as post-processing. Crucially, the prompt itself is refined by the LLM: at each iteration the model analyzes its own errors and proposes targeted edits to its rules and examples. Through 18 such self-refinement cycles, our system achieved an F1-Strict of 0.53 and F1-Relaxed of 0.60 on the test set, ranking first among all participating teams under both evaluation criteria.

1 Introduction

The opioid epidemic remains a critical public health crisis, with social media platforms serving as rich, yet underutilized sources of first-person accounts of substance use consequences (Garnett and Miniño, 2026; Sarker et al., 2020). Task 7 of the #SMM4H-HeaRD 2026 shared task (Social Media Mining for Health/Health Real-World Data Workshop Organizers, 2026; Dey et al., 2025) challenges participants to perform span-level named entity recognition (NER) on Reddit posts, identifying two entity types: **ClinicalImpacts** (physical or psychological consequences such as overdose, withdrawal, or depression) and **SocialImpacts** (social, occupational, or relational consequences such as incarceration, job loss, or homelessness).

This task presents several challenges that make it particularly difficult for traditional NER approaches. The text is informal, containing slang (e.g., "fell out" for overdose, "kicked" for withdrawal), abbreviations ("WD", "PAWS"), with highly variable phrasing. Entity boundaries are often ambiguous, especially for multi-word expressions like "arrested for multiple felonies" or "lost everything I loved in life." Furthermore, the task requires distinguishing first-person self-reports from third-person narratives, hypothetical statements, and general observations.

Rather than fine-tuning a pretrained language model, we adopt a prompt-engineering approach using GPT-5.4 (OpenAI, 2026) with structured few-shot examples. Our key insight is that the annotation guidelines for this task encode complex decision logic that is difficult to learn from limited labeled data, but can be effectively communicated through carefully structured natural language rules and representative examples. We iteratively refine these rules through systematic error analysis over 18 prompt versions, treating the prompt itself as the "model" to be optimized. Notably, the refinement is performed *by the LLM itself*: at each iteration, the model inspects its own error reports and proposes targeted edits to its own rules and few-shot examples, forming an autonomous self-improvement loop.

2 Related Work

LLM-based NER. Recent work has explored LLMs for named entity recognition, moving beyond traditional sequence labeling. NER has been reformulated as machine reading comprehension, enabling zero-shot transfer across entity types (Li et al., 2020). LLMs have also been shown to perform NER through in-context learning with carefully designed prompts, achieving competitive results without task-specific fine-tuning (Wang et al., 2025). Our work extends this line to informal social media text with complex annotation guidelines, using structured few-shot prompting.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 28–35

Prompt Optimization. Automatic prompt engineering has emerged as an alternative to manual prompt design. APE showed that LLMs can generate and select effective prompts automatically (Zhou et al., 2023). Task-aware prompt optimization has also been proposed through iterative refinement (Agarwal et al., 2025). Recent survey work provides a broader synthesis of these techniques (Ramnath et al., 2025). Our approach differs in that we optimize structured source code, including explicit rules, ontology definitions, and categorized examples, rather than opaque prompt text. The optimization also targets a complex span-level NER task that requires fine-grained boundary decisions.

3 System Description

3.1 Overview

Our system operates as a two-stage pipeline: (1) a first-person classification pre-filter that identifies whether a text is a self-report, and (2) LLM-based NER that performs span-level entity extraction using structured few-shot prompting, followed by BIO constraint enforcement. Both stages use GPT-5.4 (OpenAI, 2026) via the OpenAI API with temperature 0.0. In practice, the first-person pre-filter was enabled in our final submission but contributed minimally as the Self-Report Test embedded in the NER prompt (Rule 1, below) already skips non-first-person entities at the span level, making the text-level filter largely redundant (see Appendix D).

3.2 Prompt Architecture

The architecture below describes the final prompt (v18) produced by the self-refinement loop (Section 3.5). The initial human-authored prompt (v1) provided only basic entity type definitions and a JSON output format, whereas the structured decision rules, boundary specifications, skip patterns, and categorized examples were all proposed by the LLM during iterative optimization and accepted by the human operator. Appendix C details the provenance of each component.

The prompt consists of a system message encoding the annotation schema and decision rules, followed by a user message containing few-shot examples and the target text. The system prompt is structured into five sections (reproduced in full in Appendix A):

I. Entity Ontology. The ontology defines `ClinicalImpacts` and `SocialImpacts` with explicit subcategories and representative terms. ClinicalImpacts include medical states (overdose, withdrawal), symptoms (nausea, depression), treatment-as-consequence (rehab, detox), and self-identification terms (addict, junkie). SocialImpacts include legal (arrest, jail), relational (divorce, estrangement), functional (job loss, homelessness), and financial consequences.

II. Core Decision Rules. Three sequential tests determine whether a candidate span should be annotated:

1. **Self-Report Test**: Did this happen to the author? Only first-person experiences are eligible.

2. **Consequence vs. Behavior Test**: Is this something that *happened to* the person (consequence) or something they *did* (behavior)? For example, "lost my job" is a consequence; "went on a bender" is a behavior.

3. **Boundary Test**: Preference for longer spans that capture the full expression, including legal phrases and descriptive modifiers.

III. Span Boundary Rules. The prompt encodes explicit include/exclude patterns for span boundaries (e.g., degree adjectives like "severe" are kept, while pure intensity words like "terrible" are stripped). See Appendix A for the complete specification.

IV. Skip Patterns. Nine categories of expressions that should *not* be annotated, including third-person references, hypothetical conditionals, general statements, and metaphorical usage.

V. Output Format. The output format specifies token indexing rules and a structured JSON format: `{"entities": [{"span", "type", "indices"}]}`.

3.3 Few-Shot Examples

We provide 65 few-shot examples organized into five cognitive categories: decision gates (17), span boundaries (19), entity type distinctions (12), special patterns (8), and error-driven additions (9). All examples are synthetic (generated by the LLM during the self-refinement loop) and verified against evaluation data using a contamination detection script. The complete set is listed in Appendix B.

3.4 BIO Constraint Enforcement

LLM outputs are parsed from JSON and converted to BIO tag sequences. A post-processing module enforces two BIO consistency constraints: orphaned `I-` tags (those without a preceding `B-` or `I-` tag of the same type) are converted to `B-`, and type mismatches where `B-X` is followed by `I-Y` trigger a new entity by converting `I-Y` to `B-Y`. Additionally, we validate that predicted span text matches the token indices (requiring $>$0.3 overlap ratio) and filter punctuation-only predictions.

3.5 Iterative Prompt Optimization

A distinguishing feature of our approach is that the prompt is refined *autonomously by the LLM itself*, rather than through manual human engineering. We implement this as a closed-loop self-refinement pipeline orchestrated by an LLM-based coding agent (Claude Code). Figure 1 illustrates the architecture. The agent harness provides the orchestrating model with tool-call access to the file system, a Python interpreter, and the evaluation scripts. At each iteration, the agent: (1) evaluates the current prompt on the training set by invoking `run_llm_ner.py` via tool call, producing structured error reports that categorize each mistake as a false negative, false positive, boundary error, or type mismatch; (2) analyzes error patterns by reading the resulting JSON error files, identifying recurring failure modes such as missed slang terms, over-broad span boundaries, or misclassified entity types; (3) edits its own prompt by issuing targeted file-edit tool calls to directly modify the system prompt rules, span boundary specifications, skip patterns, and few-shot examples in `src/llm/llm_ner.py`; and (4) validates the modified prompt on the development set, accepting changes only if metrics improve and rejecting them otherwise.

Each iteration is also validated against a contamination detection script that checks for textual overlap between few-shot examples and evaluation data. This process is automated, where the human role is limited to initiating the loop and reviewing the final output, while all error analysis, rule formulation, example generation, and code edits are performed by the model.

4 Results

4.1 Development Set Performance

Table 1 presents the development set results for our final system (v18). SocialImpacts (0.74) slightly outperforms ClinicalImpacts (0.71), which we attribute to availability and phrasing of social consequence examples in the training set.

System / Entity	P	R	F1-R
Overall	0.75	0.69	0.72
ClinicalImpacts	—	—	0.71
SocialImpacts	—	—	0.74

Table 1: Development set relaxed metrics (v18). F1-R = relaxed F1, P = precision, R = recall. F1-Strict for our system: 0.61.

4.2 Test Set Performance

Table 2 presents the official test set results. Our system achieved an F1-Strict of 0.53 and F1-Relaxed of 0.60, placing first among all participating teams under both evaluation criteria. For reference, Dey et al. (2025) report a fine-tuned DeBERTa-large (He et al., 2021) achieving 0.61 relaxed F1 on a held-out test split of the same underlying dataset, suggesting that our zero-training approach performs comparably to supervised fine-tuning (0.60 vs. 0.61).

	F1-Strict	F1-Relaxed
DeBERTa-large†	—	0.61
Team Gazoo!	0.53	0.60
All teams (mean)	0.46	0.55
All teams (median)	0.48	0.58

Table 2: Test set results. Our system ranked first among all participating teams under both criteria. †DeBERTa-large from Dey et al. (2025), evaluated on their held-out test split of the same dataset; the test splits differ, so comparison is approximate.

4.3 Prompt Optimization Trajectory

Over 18 iterations, our self-refinement process improved relaxed F1 from 0.41 (v1) to 0.72 (v18) on the development set. The largest gains came in early iterations (v1–v6, +22 points) from introducing the entity ontology and decision rules. Performance plateaued at approximately 65 few-shot

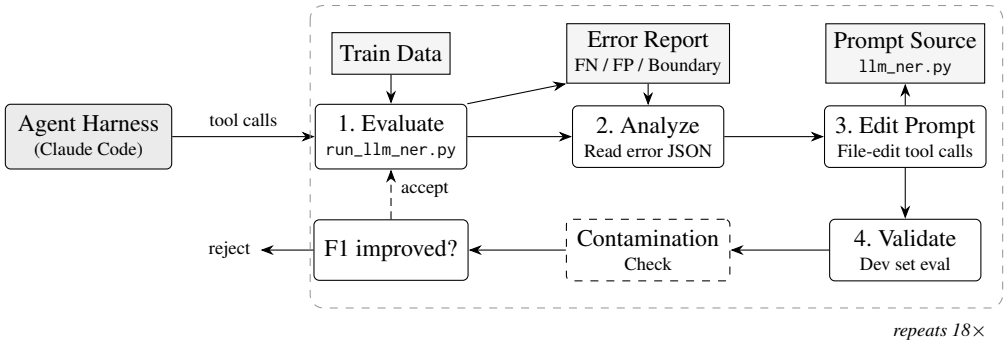

Figure 1: Self-refinement loop. The agent harness (Claude Code) orchestrates each iteration via tool calls: it runs evaluation on the training set, reads the resulting error reports, and issues file-edit tool calls to modify the prompt source code. The updated prompt is then validated on the development set with a contamination check. Changes are accepted only if F1 improves; otherwise they are reverted.

examples; adding further examples degraded performance, suggesting a ceiling imposed by prompt length and in-context learning capacity. Table 3 traces selected versions.

Ver.	F1-S	F1-R	Key Change
v1	0.23	0.41	Initial prompt
v6	0.52	0.63	Entity ontology + decision rules
v12	0.59	0.66	SocialImpacts examples
v18	0.61	0.72	Cognitive example categories

Table 3: Optimization trajectory on the development set. F1-S = F1-Strict; F1-R = F1-Relaxed. All changes were proposed and implemented through the LLM self-refinement loop.

5 Discussion

Dev-to-Test Drop. We observe a notable drop from development (0.72 relaxed F1) to test (0.60 relaxed F1). This 12-point gap likely reflects overfitting to the development distribution, which is a known risk of iterative prompt optimization on a fixed evaluation set (Zhou et al., 2023; Agarwal et al., 2025). Each iteration reduces development set error but may encode spurious correlations specific to that sample. We mitigate this partially through contamination detection (ensuring few-shot examples do not overlap evaluation data) and accept/reject gating (changes must improve metrics to be retained). However, neither safeguard prevents the rules themselves from overspecializing to development set patterns. Future work could address this through held-out validation splits, cross-validation across prompt versions, or regularization strategies such as capping the total number of rules.

Self-Refinement Dynamics. We observed diminishing returns after approximately 12 iterations, with the error distribution shifting from false negatives (missing entity categories) in early versions to boundary disagreements in later versions. A key strength of this approach is full interpretability. Every annotation decision traces back to explicit rules and examples in the prompt, enabling rapid adaptation without retraining.

Limitations

The approach relies on a proprietary LLM (GPT-5.4) via a commercial API, limiting reproducibility. The 12-point dev-to-test gap indicates that further work is needed to improve robustness to distribution shift in prompt-based NER systems. Additionally, the system was developed and evaluated solely on English-language Reddit text, and its generalizability to other platforms, languages, or substance use domains has not been assessed.

Ethics Statement

This work involves processing social media text describing substance use and its consequences, which may contain sensitive personal disclosures. All data was provided by the shared task organizers and used solely for research purposes.

Reproducibility

The complete system prompt and all 65 few-shot examples are reproduced in Appendices A and B.

References

Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, and Akshay Nambi. 2025. PromptWizard: Optimizing prompts via task-aware, feedback-driven self-evolution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19974–20003, Vienna, Austria. Association for Computational Linguistics.

Sumon Kanti Dey, Jeanne M. Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. *Preprint*, arXiv:2508.19467.

Matthew F. Garnett and Arialdi M. Miniño. 2026. Drug overdose deaths in the united states, 2023–2024. NCHS Data Brief 549, National Center for Health Statistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.4. Accessed: 2026-05-28.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, and 2 others. 2025. A systematic survey of automatic prompt optimization techniques. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33078–33110, Suzhou, China. Association for Computational Linguistics.

Abeed Sarker, Annika DeRoos, and Jeanmarie Perrone. 2020. Mining social media for prescription medication abuse monitoring: A review and proposal for a data-centric framework. *Journal of the American Medical Informatics Association*, 27(2):315–329.

Social Media Mining for Health/Health Real-World Data Workshop Organizers. 2026. Social media mining for health/health real-world data (#smm4h-heard) 2026 workshop and shared tasks: Task 7 – extraction of social and clinical impacts of substance use from social media posts. Accessed: 2026-05-28.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Complete System Prompt

The following is the full system prompt (v18) passed to GPT-5.4 at inference time. This prompt was produced through 18 iterations of the self-refinement loop described in Section 3.5.

```
You are an expert NER system for detecting
self-reported substance use impacts in social media.

## ENTITY ONTOLOGY

**ClinicalImpacts** = Consequences TO the body/mind:
- Medical states: overdose, withdrawal, addiction,
  relapse, dependence, tolerance, cravings
- Symptoms: physical (nausea, pain, tremors,
  hairloss, acne), psychological (depression,
  anxiety, psychosis), informal ("felt terrible",
  "issues")
- Treatment-as-consequence: rehab, detox,
  hospitalization, therapy, MAT, clinic, doctor
  (when treatment received)
- Self-identification: addict, junkie, habit
  (standalone, NOT "[substance] habit")
- Sensitive (MUST annotate): suicidal ideation,
  overdose, died/fell out (strip "almost/nearly"
  per boundary rules)
- Slang/abbreviations: "kicked", "tweaking" (as
  past consequence), "fell out", "wds", "WD", "PAWS"
- Vague terms in context: "issues", "problems"
  when describing symptoms

**SocialImpacts** = Consequences TO life circumstances:
- Legal: arrest, jail, prison, probation, charges,
  felonies - INCLUDE charge details ("arrested for
  multiple felonies")
- Relational: divorce, estrangement, custody loss,
  "family disowned", "broke up", "bad relationship"
- Functional: job loss, homelessness, eviction,
  "lost everything", "unable to function"
- Financial: debt, "had no money", "dropping $100s",
  "destroys your wallet"
- Recovery programs: AA, NA (even when mentioned in
  others' context if author participates)
- Violence/abuse: include FULL phrase ("grabbed me
  by the throat", "physical assault on a helpless
  man")
- Life crisis: "rock bottom"
- Illegal activities: "dealing drugs", "stealing"
- Family threats: "threaten to take away my car"
- Extreme loss: "lost everything...other than
  [exception]" - include the exception clause

---

## CORE DECISION RULES

**Rule 1: SELF-REPORT TEST**
Ask: "Did THIS happen TO the author (or is author
describing their own state)?"
- YES -> candidate for annotation
- NO (purely others' experience with no author
  involvement) -> SKIP
- NOTE: "your" addressing the author = self-report
  ("destroys your wallet" in "Knowing heroin
  destroys your wallet")

**Rule 2: CONSEQUENCE vs BEHAVIOR TEST**
Ask: "Is this what HAPPENED TO them, or what they
DID?"
- Consequence (result) -> annotate
```

```
- Behavior (action) -> SKIP: "bender", "using"
- EXCEPTION: "relapse" as noun = annotate;
  "relapsed" as verb = context-dependent
- EXCEPTION: "habit" standalone = annotate;
  "[substance] habit" = SKIP

**Rule 3: BOUNDARY TEST - PREFER LONGER SPANS**
When in doubt, include MORE context rather than less:
- Include full legal phrases: "arrested for multiple
  felonies" NOT just "arrested"
- Include full loss phrases: "lost everything I loved
  in life" NOT just "lost everything"
- Include descriptive extent: "strong cravings",
  "severe cravings" (KEEP the adjective)
- Include full action descriptions: "failed to pay
  on traffic tickets" NOT just "traffic tickets"

---

## SPAN BOUNDARY RULES

**INCLUDE** (keep these in span):
- DEGREE ADJECTIVES: strong|severe|high|massive|
  chronic + symptom -> KEEP
- [VERB + FULL OBJECT]: "arrested for multiple
  felonies", "lost everything I loved"
- [ACTION + STATE]: "wake up sick", "pass out cold"
- [TYPE-SPECIFIER + CORE]: "meth induced psychosis",
  "benzodiazepine dependence"
- [COMPOUND TREATMENT]: "detox / rehab", "inpatient
  psych" (single entity)
- [FULL INABILITY]: "couldn't make it to work",
  "unable to function"
- [NEGATED-BUT-ACKNOWLEDGED]: "didn't lose my
  apartment", "didn't fuck my life up"

**EXCLUDE** (strip from spans):
- PURE INTENSITY (no meaning): horrible|terrible|
  awful|mild -> strip
  - BUT KEEP: strong|severe|high|massive|chronic
- ONSET VERB: started|began|developed -> strip
- MOVEMENT VERB: "went to"|"checked into" -> strip
- "got" + SLANG: "got dope sick" -> "dope sick"
- POSSESSIVE: "my" before body parts -> strip
- TEMPORAL SUFFIX: "every day", "in the shower"
  at END of symptom -> strip
- NEAR-MISS MODIFIERS: "almost/nearly" + [overdosed|
  died] -> keep only the event
- INTENSITY MODIFIERS: "full on" + symptom -> keep
  only the symptom

---

## SKIP PATTERNS (DO NOT ANNOTATE)

| Pattern | Reason |
|---------|--------|
| Others' experience | not self-report |
| Hypothetical conditional | not event |
| General statement | not personal |
| Metaphorical only | figurative |
| Coping/managing context | not experiencing |
| Emergency scene incidentals | not consequence |
| Emotions alone | not consequence |
| Positive recovery | achievement |
| Indirect consequence | not from substance |

---

## SPECIAL CONTEXTS

**submission_title**:
- Describing OWN symptom (even with "?") -> ANNOTATE
- Asking about OTHERS ("Anyone...?") -> SKIP

**Parenthetical content**: ALWAYS check inside ()

**"Knowing X destroys your Y" pattern**: "your" =
self-reference -> ANNOTATE

---

## INDEX RULES
- Tokens indexed as [0], [1], [2]...
- Punctuation has own index - NEVER include in spans
- Indices must match tokens exactly

## OUTPUT FORMAT
{"entities": [{"span": "text",
  "type": "ClinicalImpacts|SocialImpacts",
  "indices": [0, 1, 2]}]}
If no entities: {"entities": []}
```

B Few-Shot Examples

All 65 few-shot examples are listed below, organized by cognitive category. Each example shows the input text and expected entity annotations. All examples are synthetic (generated during the self-refinement loop) and verified for no overlap with evaluation data.

Category A: Decision Gates (17 examples)

1. **Self-report + multiple entities (PASS):** "I ended up unemployed with chronic insomnia." → *unemployed* [Social], *chronic insomnia* [Clinical]

2. **Others' experience (SKIP):** "My brother went through terrible withdrawals." → { }

3. **Behavior vs consequence:** "After the bender, I lost my license." → *lost my license* [Social]

4. **[substance] habit (SKIP):** "I developed a terrible oxy habit back then." → { }

5. **Standalone habit (PASS):** "I had the same exact habit as him." → *habit* [Clinical]

6. **Hypothetical "could have" (SKIP):** "I could have easily died that night." → { }

7. **"almost died" – annotate core:** "I almost died four times." → *died* [Clinical]

8. **General statement (SKIP):** "Benzo withdrawal is a nightmare for most people." → { }

9. **Personal context (PASS):** "The withdrawal was an absolute nightmare." → *nightmare* [Clinical]

10. **Metaphorical "hell" (SKIP):** "Using daily was absolute hell for me." → { }

11. **Location alone (SKIP); consequence (PASS):** "I ended up at the ER with an IV in my arm." → *IV in my arm* [Clinical]

12. **Emergency incidentals (SKIP):** "Woke up in the hospital, car impounded, wallet stolen." → { }

13. **Positive recovery (SKIP):** "Today I have a job and an apartment." → { }

14. **Background info (SKIP):** "I have an addiction history that complicates things." → { }

15. **Coping/managing (SKIP):** "That really helps control cravings for me." → { }

16. **Hypothetical conditional (SKIP):** "I would use until I ran out of money." → { }

17. **Indirect consequence (SKIP):** "I told them I was in recovery so I was not getting hired." → { }

Category B: Span Boundaries (19 examples)

1. **KEEP degree adj:** "I found myself with strong cravings at night." → *strong cravings* [Clinical]

2. **Strip intensity adj:** "I had terrible nightmares every night." → *nightmares* [Clinical]

3. **Full legal phrase:** "I've been arrested for multiple felonies several times." → *arrested for multiple felonies* [Social]

4. **Full loss phrase:** "I literally lost everything I loved in life." → *lost everything I loved in life* [Social]

5. **Full action phrase:** "I was keeping up but failed to pay on traffic tickets." → *failed to pay on traffic tickets* [Social]

6. **Strip "got" before slang:** "When I ran out I got dope sick fast." → *dope sick* [Clinical]

7. **Strip "got" before "caught":** "I used until I got caught by a DT." → *caught by a DT* [Social]

8. **KEEP "became hooked":** "I became hooked on benzos fast." → *became hooked* [Clinical]

9. **Strip causal suffix:** "I was dizzy and nauseous from the pills." → *dizzy and nauseous* [Clinical]

10. **Strip temporal suffix:** "I lose clumps in the shower every single day." → *lose clumps* [Clinical]

11. **Strip duration prefix:** "I completed a ninety day program." → *program* [Clinical]

12. **Strip possessive:** "My hands were shaking uncontrollably." → *hands were shaking uncontrollably* [Clinical]

13. **KEEP type-specifier:** "The doctors diagnosed alcohol induced neuropathy." → *alcohol induced neuropathy* [Clinical]

14. **KEEP verb + object:** "The drugs destroyed my liver." → *destroyed my liver* [Clinical]

15. **KEEP full symptom:** "I could hardly get the words out without getting sick." → *hardly get the words out without getting sick* [Clinical]

16. **Compound treatment:** "They put me in a detox / rehab facility." → *detox / rehab* [Clinical]

17. **Negated-but-acknowledged:** "I didn't fuck my life up as bad as some." → *didn't fuck my life up as bad as some* [Social]

18. **Strip "me" after relationship verb:** "My parents disowned me after the arrest." → *parents disowned* [Social], *arrest* [Social]

19. **Strip preposition:** "I spent six months in jail for possession." → *jail* [Social]

Category C: Entity Type Distinctions (12 examples)

1. **Treatment facility = Clinical:** "They admitted me to the detox unit immediately." → *detox unit* [Clinical]

2. **AA/NA = Social:** "I meet people in AA who got sober with rehab." → *AA* [Social]

3. **Legal = Social (full phrase):** "They charged me with possession and intent to distribute." → *charged me with possession and intent to distribute* [Social]

4. **Relationship = Social:** "We broke up because of my using." → *broke up* [Social]

5. **Violence = Social (full phrase):** "My father grabbed me by the throat." → *father grabbed me by the throat* [Social]

6. **"relapse" as noun = Clinical:** "The first time since my last relapse ended." → *relapse* [Clinical]

7. **Illegal activity = Social:** "I found money for drugs by dealing drugs." → *dealing drugs* [Social]

8. **Family threats = Social:** "They threaten to take away my car if I don't comply." → *threaten to take away my car* [Social]

9. **Rock bottom = Social:** "I hit rock bottom and finally got help." → *rock bottom* [Social]

10. **Slang "kicked" = Clinical:** "Kicked at 170, was sick for six months." → *Kicked at 170* [Clinical], *sick* [Clinical]

11. **"call EMS" = Clinical:** "They had to call EMS because I collapsed." → *call EMS* [Clinical], *collapsed* [Clinical]

12. **"fell out" = Clinical:** "I shot up and almost fell out." → *almost fell out* [Clinical]

Category D: Special Patterns (8 examples)

1. **Parenthetical content:** "Felonies (violent felonies from a bad relationship and substance abuse)." → *Felonies* [Social], *violent felonies* [Social], *bad relationship* [Social], *substance abuse* [Clinical]

2. **"your" = self-reference:** "Knowing heroin destroys your wallet and family, I still use." → *destroys your wallet and family* [Social]

3. **Sensitive content:** "I thought about killing myself every day." → *killing myself* [Clinical]

4. **submission_title own symptom:** "submission_title: Small, pinprick red dots on skin?" → *Small, pinprick red dots on skin* [Clinical]

5. **submission_title asking others (SKIP):** "submission_title: Any supplements to reduce acne?" → {}

6. **Slang abbreviation:** "These wds are killing me." → *wds* [Clinical]

7. **Doctor in treatment context:** "I was shocked when my second doctor didn't do induction." → *doctor* [Clinical]

8. **Full financial phrase:** "Not having enough money to even be well was eye opening." → *Not having enough money to even be well* [Social]

Category E: Error-Driven Additions (9 examples)

1. **Vague "issues" in context:** "I tried it and had issues at the end." → *issues* [Clinical]
2. **[substance] habit still SKIP:** "I had the same kratom habit as him." → {}
3. **Standalone symptoms:** "The worst ones are hairloss and insomnia." → *hairloss* [Clinical], *insomnia* [Clinical]
4. **"tweaking" as past consequence:** "That was my last time tweaking." → *tweaking* [Clinical]
5. **KEEP "mental cravings":** "The mental cravings decreased by 80%." → *mental cravings* [Clinical]
6. **"WD" abbreviation:** "Would have used this during my WD." → *WD* [Clinical]
7. **"addiction" standalone:** "My only addiction is opiates." → *addiction* [Clinical]
8. **submission_title with context:** "submission_title: My most recent trip to jail." → *My most recent trip to jail* [Social]
9. **Full inability phrase:** "I couldn't get to the clinic." → *couldn't get to the clinic* [Social]

Note: One additional example (E10: "Job, house, friends all gone because of addiction" → full phrase [Social]) is included in the system but omitted here for space.

C Human vs. LLM Provenance

Table 4 clarifies which prompt components were part of the initial human-authored prompt (v1) versus those introduced through the LLM self-refinement loop.

Component	v1 (Human)	v18 (LLM-refined)
Entity Ontology	Type names + 3–4 examples each	Detailed subcategories, slang, abbreviations
Decision Rules	None	3 sequential tests
Span Boundary	None	8 INCLUDE + 8 EXCLUDE patterns
Skip Patterns	"Only annotate first-person"	9 explicit categories
Few-Shot Examples	5 generic examples	65 across 5 cognitive categories
Output Format	JSON spec	Same (unchanged)

Table 4: Provenance of prompt components. The human authored the initial skeleton (v1); all structured rules and categorized examples were proposed by the LLM and accepted by the human operator during the self-refinement loop.

D First-Person Filtering Stage

Our pipeline includes an optional text-level first-person classification stage that runs before NER. This binary classifier determines whether an input text constitutes a first-person self-report about substance use. Texts classified as non-first-person are assigned empty entity annotations without invoking the NER prompt.

This stage was enabled in our final submission. However, its contribution is largely redundant with Rule 1 (Self-Report Test) in the NER prompt itself: the NER system already applies first-person filtering at the span level through its skip patterns ("Others' experience" and "General statement" categories). The text-level pre-filter operates at a coarser granularity and cannot handle mixed-perspective texts where some spans are first-person and others are not.

In ablation testing on the development set, disabling the pre-filter resulted in less than 0.5 F1 change, confirming its minimal independent contribution. We retain it as a configurable pipeline component that may benefit efficiency in deployment scenarios with high volumes of clearly non-relevant text.