Modulating Multi-Label Tendency in Zero-Shot LLM Coding: The Effect of Output Structure on CDSS Feedback Analysis

Hyunwoo Choo
AITRICS
hwchoo@aitrics.com

Sungsoo Hong
AITRICS
sshong@aitrics.com

Abstract

Large language models (LLMs) often default to single-label classification in zero-shot multi-label tasks—a tendency we term **conservative default**. While few-shot prompting mitigates this, it introduces **example bias**. We evaluate zero-shot strategies to modulate this tendency using 1,441 healthcare feedback records and two LLMs. We compare instruction-based methods with structural constraints that modify the token generation sequence, specifically an **Enumeration-First (Enum-First)** format requiring explicit domain enumeration before selection. Results show that structural constraints substantially reduce single-label rates (Magistral: 96% → 19%; Qwen3: 54% → 0.0%), though the latter suggests potential overcorrection compared to human baselines (16.7–41.3%). These findings indicate that while output structure is a potent modulator of classification behavior by shifting the decision point upstream, its effect magnitude is model-dependent, necessitating empirical calibration to prevent spurious associations.

1 Introduction

Large language models (LLMs) have demonstrated capability in clinical knowledge encoding (Singhal et al., 2023) and information extraction (Agrawal et al., 2022), emerging as promising tools for qualitative coding (Xiao et al., 2023). While prompting strategies like Chain-of-Thought (Wei et al., 2022) and Self-Consistency (Wang et al., 2023) have improved reasoning, their application to **multi-label qualitative coding** remains understudied. Zero-shot approaches are particularly advantageous as they require no labeled training data.

However, qualitative research often employs simultaneous coding (Saldaña, 2013), where a single text segment is assigned multiple codes when it addresses multiple themes. For instance, user feedback stating *"The alarm is too frequent, disrupting patient care"* might be coded as both a technology issue (alarm frequency) and an organizational issue (workflow disruption).

The Few-Shot Dilemma. A natural solution to improve LLM classification is few-shot prompting (Brown et al., 2020). However, for simultaneous coding with multi-label outputs, few-shot approaches face fundamental challenges:

- **Combinatorial explosion:** With 7 Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) domains, there are $7 \times 6 = 42$ possible dominant-associated pairs. Providing examples for all combinations is impractical.
- **Example bias:** Providing a subset of examples biases the model toward those specific combinations. In pilot experiments, we observed that examples with "Technology + Organization" pairs led the model to over-assign this combination.

This creates a dilemma: few-shot improves performance but introduces systematic bias toward the provided examples. We therefore sought zero-shot strategies that could modulate multi-label tendency without content-specific examples.

Feedback: "I'm often out for rounds and prefer mobile access to the screening list. The hospital IT team and electronic medical record (EMR) vendor have been uncooperative."

Baseline (Standard Zero-shot):
▷ Dominant: `TECHNOLOGY` (mobile feature request)
▷ Associated: `NONE` (Missed organizational barrier)

Ours (Enum-First):
▷ All Relevant: `[TECHNOLOGY, ORGANIZATION]`
▷ Dominant: `TECHNOLOGY`
▷ Associated: `ORGANIZATION` (IT non-cooperation)

Table 1: Comparison of standard vs. Enum-First output format. The baseline captures the feature request but misses the organizational barrier; Enum-First captures both.

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 172–179
March 28, 2026 ©2026 Association for Computational Linguistics

The Conservative Default. This study investigates a phenomenon we term conservative default, characterized by a high frequency of single-label assignments in zero-shot multi-label tasks. We hypothesize that the sequential nature of standard classification prompts—where the secondary domain is requested after the primary—may lead models to favor a null assignment ("NONE") for the second slot despite the presence of multi-themed content (Table 1).

To test this, we applied zero-shot prompts to classify 1,441 feedback records from a clinical decision support system (CDSS) using the NASSS framework (Greenhalgh et al., 2017). The baseline prompt instructed the models to assign a dominant domain and an optional associated domain. In this condition, the models assigned no associated domain in 96% (Magistral) and 54% (Qwen3) of cases. We investigate whether these high single-label rates are a product of output structure rather than the underlying data distribution alone.

Research Questions. Specifically, this study addresses three questions:

1. Can prompt engineering modulate single-label tendency in zero-shot multi-label classification without introducing example bias?
2. Which prompting strategies are most effective, and does effectiveness vary across models?
3. What are the potential risks of over-correction?

2 Methods

2.1 Dataset

We used 1,441 free-text feedback records from users of a clinical decision support system (CDSS) deployed across 166 hospitals in South Korea (2022–2025). Feedback was collected through free-form submission and included complaints, suggestions, and questions about the system.

2.2 Coding Framework

We applied the NASSS framework (Greenhalgh et al., 2017), which categorizes implementation complexity into seven domains (e.g., Technology, Organization, Adopter System).

Each feedback item was coded for a dominant domain (primary issue) and an associated domain (secondary domain, or "NONE").

2.3 Prompting Strategies

We designed four strategies targeting distinct levels of prompting intervention (Table 2; full prompt details in Appendix A). Strategies A–C represent instruction-based approaches at three levels: **lexical** (surface-form cues), **schematic** (abstract relational patterns), and **procedural** (task decomposition), respectively. Strategy D operates orthogonally to these by modifying not the instruction content but the **output structure** itself, allowing it to be combined with any instruction-based strategy. Specifically, Strategy A provides linguistic triggers (e.g., causal and contrastive markers) to identify cross-domain links. Strategy B provides domain-agnostic abstract patterns to avoid content-specific bias. Strategy C introduces a two-step reasoning process that enumerates domains before assigning roles. Strategy D (**Enum-First**, for "enumeration before selection") enforces a structural constraint by requiring a list of all relevant domains before selecting the dominant one.

Strategy	**Core Logic / Instruction**
A: Linguistic	Identify multi-domain grammatical markers (e.g., causal, contrastive markers).
B: Slot-Based	Use domain-agnostic abstract patterns (e.g., "[X] causes [Y]") to map relations.
C: Two-Step	Procedural instruction: (1) Enumerate all domains → (2) Assign dominant/associated.
D: Enum-First	Modified output schema requiring `all_relevant_domains` list before selection.

Table 2: Summary of tested zero-shot prompting strategies.

2.4 Experimental Conditions

We tested configurations across two dimensions: (1) prompting instructions (A, B, C) and (2) output format (standard vs. Enum-First). See Table 3 for configurations.

2.5 Models and Evaluation

To assess generalizability, we evaluated two LLMs with different characteristics:

- **Magistral-Small-2509**: A reasoning-focused model.
- **Qwen3-30B-A3B-Instruct-2507**: A larger instruction-tuned model.

Both models were run via vLLM (Kwon et al., 2023) with guided JSON decoding (temperature = 0). Because the goal of this study is to charac-

Configuration	A	B	C	D
Baseline				
+Linguistic (A)	✓			
+SlotPatterns (B)		✓		
+TwoStep (C)			✓	
+A+C	✓		✓	
+B+C		✓	✓	
+A+B+C	✓	✓	✓	
Enum-First (D)				✓
Enum-First+A+C	✓		✓	✓

Table 3: Experimental configurations. A = Linguistic Markers, B = Slot-Based Patterns, C = Two-Step Listing, D = Enum-First Output.

terize how output structure modulates multi-label tendency—not to evaluate correctness against a gold standard—we report behavioral metrics rather than classification accuracy: (1) NONE rate—the percentage of responses with no associated domain, indicating single-label classification; (2) domain distribution stability; and (3) enumeration depth under the Enum-First format.

3 Results

3.1 Effect of Structural Constraints

Table 4 presents the effect of each strategy on single-label classification rates (NONE rate) across both models.

Configuration	Magistral	Qwen3
Baseline	96.0%	54.1%
+Linguistic (A)	90.1%	42.1%
+SlotPatterns (B)	92.6%	48.4%
+TwoStep (C)	83.8%	25.1%
+A+C	79.8%	22.6%
+B+C	82.8%	26.9%
+A+B+C	78.0%	29.7%
Enum-First (D)	39.9%	0.6%
Enum-First+A+C	19.1%	0.0%

Table 4: NONE rate (single-label %) by configuration. Both models show reductions, with Enum-First producing the largest effects.

Both models responded to all strategies, but the Enum-First format produced the largest reductions. Effect magnitude was model-dependent: Qwen3 reached 0% while Magistral stabilized at 19%.

3.2 Number of Domains Listed

Table 5 shows the distribution of relevant domains identified under the Enum-First format.

	Magistral		Qwen3	
# Domains	D	D+A+C	D	D+A+C
1 (single)	39.9%	19.1%	0.6%	0.0%
2 (dual)	59.6%	79.0%	86.0%	71.8%
3+ (triple)	0.5%	1.9%	13.4%	28.2%

Table 5: Distribution of domain counts under Enum-First format. D = Enum-First alone; D+A+C = Enum-First with Strategies A and C.

3.3 Domain Distribution Patterns

Dominant domain distributions remained stable across configurations (Technology: ∼77–85%), indicating that primary classification was preserved. Table 6 shows secondary domain distributions; Organization emerged as the primary associated domain under Enum-First for both models.

	Magistral		Qwen3	
Assoc.	+AC	Enum	+AC	Enum
Organization	31.8%	65.0%	42.2%	46.7%
Adopter	51.9%	12.7%	24.7%	17.8%
Value Prop	1.7%	7.3%	5.9%	11.9%

Table 6: Secondary domain distribution (excl. NONE). Top 3 domains shown. +AC = instruction-based (+A+C); Enum = Enum-First+A+C.

4 Discussion

4.1 Effect of Output Structure

The results indicate that output structure influences multi-label tendency. Both models responded to the Enum-First format by producing more multi-label outputs, supporting the hypothesis that requiring a list before selection structurally prevents single-label shortcuts. This aligns with recent findings that grammar-constrained decoding influences structured task performance (Geng et al., 2023) and the established perspective of treating multi-label classification as a sequence generation problem (Yang et al., 2018). The Enum-First format functions similarly to approaches that separate content generation from output structuring (Li et al., 2024), reducing the burden of handling orthogonal subtasks simultaneously.

Three factors may underlie this effect: (1) the model cannot bypass enumeration when the schema requires a list, (2) it must evaluate domains before prioritizing, and (3) access to the single-label default is structurally reduced. In autoregressive terms, the standard format requests the associated domain only after the dominant domain has been committed to the context window. At this point, the model's hidden states may be conditioned on a single-label trajectory, making "NONE" a low-entropy default that satisfies the schema with minimal additional computation. The Enum-First format shifts this decision upstream: by requiring the population of `all_relevant_domains` first, the model evaluates each domain's relevance before any prioritization occurs. Consequently, the multi-label status is determined by the length of the generated list rather than as a secondary reconsidered step, effectively neutralizing the structural bias toward single-label shortcuts.

The stability of dominant domain assignments (Section 3.3) indicates that this modulation affects secondary domain assignment without distorting primary classification. Notably, because the Enum-First intervention operates at the output schema level rather than through language- or domain-specific instructions, it is in principle applicable to any multi-label task where structured decoding is available, regardless of language, taxonomy, or application domain.

4.2 The Risk of Over-Correction

The reduction of Qwen3's NONE rate to 0.0% under the Enum-First format indicates a potential over-correction effect. While the same format brought Magistral's NONE rate from 96.0% to 19.1%, the near-elimination in Qwen3 suggests that strong structural constraints may force models to generate spurious associations when they already exhibit moderate multi-label tendencies. Qualitative inspection supports this: minimal feedback such as "XAI" (3 characters) or "메모란" (memo field; 3 characters) received associated domains (Value Proposition and Adopter System, respectively) under Enum-First, despite being single-topic feature requests where no cross-domain link is evident. Preliminary human annotation on a subset (n=150) shows trends consistent with this interpretation: annotators identified single-domain feedback in 16.7–41.3% of cases, suggesting that a near-zero NONE rate is implausible. These findings indicate that the optimal degree of structural intervention is model-dependent, and practitioners should calibrate output schemas against human-validated samples.

4.3 The Limitations of Abstract Patterns

Strategy B, which provided domain-agnostic patterns (e.g., "[X] leads to [Y]"), achieved only marginal improvement. We attribute this to the gap between the rigid simplicity of the provided patterns and the high linguistic complexity of healthcare feedback. Real-world feedback often embeds multi-domain associations through nuanced context that does not map directly onto fixed causal or contextual slots. Furthermore, unlike the Enum-First format which modifies the actual output schema, Strategy B relies solely on text-based instructions. Our findings suggest that for zero-shot multi-label tasks, the sequence of token generation (enumeration before selection) is a more potent modulator of classification behavior than abstract procedural guidance.

4.4 Zero-Shot vs. Few-Shot Trade-off

The Enum-First approach achieved multi-label modulation without examples, avoiding the example bias and combinatorial scalability issues discussed in Section 1. Preliminary few-shot experiments confirmed this trade-off (Appendix B).

5 Conclusion

This study addressed three questions regarding zero-shot multi-label classification with LLMs.

RQ1 (Modulation without example bias): The Enum-First output format reduced single-label rates without requiring content-specific examples, offering a path around the combinatorial and bias issues inherent to few-shot prompting.

RQ2 (Effectiveness and model variation): Structural constraints on output format showed larger effects than instruction-based strategies alone. However, effect magnitude was model-dependent: a format that achieved reasonable calibration in one model produced near-total elimination of single-label outputs in another.

RQ3 (Over-correction risk): The near-zero NONE rate observed in Qwen3 suggests that strong structural constraints may induce spurious associations in models with moderate baseline tendencies. This highlights the need for validation against human judgment.

For practitioners, output schema modification offers a useful lever to modulate LLM classification behavior. However, the same structural constraint that corrects under-classification in one model may induce over-classification in another; the appropriate degree of intervention should be calibrated empirically based on baseline model behavior.

Limitations

This study has several limitations. First, we evaluated only two LLMs; results may vary across other models. Second, Strategy A relies on Korean-specific grammatical markers (e.g., 때문에 'because of', 지만 'but') that would require adaptation for other languages; generalizability of the Enum-First format is discussed in Section 4.1. Third, while preliminary human annotation shows consistent trends, further validation with the complete dataset is ongoing. Finally, we did not compare our approach against fine-tuning methods, as our focus was on maximizing the utility of off-the-shelf LLMs without the resource overhead of parameter updates.

Ethical Statement

This study used de-identified system-usability feedback from healthcare professionals; no patient health information was involved. As the data originated from routine product-operations channels with no identifiable patient data, formal Institutional Review Board (IRB) approval was not required under applicable guidelines. All personally identifiable information (user names, hospital names, contact details) was removed prior to analysis in accordance with applicable data-protection regulations.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952.

Trisha Greenhalgh, Joseph Wherton, Chrysanthi Papoutsi, Jennifer Lynch, Gemma Hughes, Christine A'Court, Susan Hinder, Nick Fahy, Rob Procter, and Sara Shaw. 2017. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, 19(11):e367.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024. A simple but effective approach to improve structured language model output for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5133–5148, Miami, Florida, USA. Association for Computational Linguistics.

Johnny Saldaña. 2013. *The Coding Manual for Qualitative Researchers*, 2nd edition. SAGE Publications.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

A Prompting Strategy Details

This appendix provides the key components of each prompting strategy evaluated in the main paper.

A.1 Strategy A: Linguistic Markers

Strategy A provides explicit Korean grammatical patterns indicating multi-domain content:

- **Causal** (“because of”, “so”; 때문에, 라서, 해서): X causes Y → code both
- **Contrast** (“but”, “although”; 인데, 지만, 는데): problem + context → code both
- **Compound** (“also”, “and”, “while”; 도, 고, 면서): simultaneous issues → code both
- **Conditional** (“in order to”, “must”; 하려면, 해야): goal + barrier → code both
- **Inability** (“cannot”; 못한다): consider underlying cause domain

Decision Rule: If any marker connects concepts from two NASSS domains, assign an associated domain. Use “NONE” only for genuinely isolated concerns.

A.2 Strategy B: Slot-Based Abstract Patterns

Strategy B provides domain-agnostic structural patterns. The [DOMAIN_X] and [DOMAIN_Y] slots can be filled by any two different NASSS domains:

```
Pattern 1: CAUSAL ("X leads to Y")
Structure: "[DOMAIN_X issue] causes [DOMAIN_Y
     consequence]"
Logic: Problem in X directly causes problem in Y
Coding: Dominant = X (source), Associated = Y (
     effect)

Pattern 2: CONTEXTUAL ("X is worse because of Y
     ")
Structure: "[DOMAIN_X problem] is exacerbated by
      [DOMAIN_Y]"
Logic: Problem in X is worsened or
     contextualized by Y
Coding: Dominant = X (main issue), Associated =
     Y (context)

Pattern 3: COMPOUND ("Both X and Y")
Structure: "[DOMAIN_X issue] and [DOMAIN_Y issue
     ]"
Logic: Two distinct issues mentioned together
Coding: Dominant = more prominent, Associated =
     other

Pattern 4: BARRIER ("Want X but Y prevents it")
Structure: "[DOMAIN_X goal] but [DOMAIN_Y
     barrier]"
Logic: Goal in one domain blocked by issue in
     another
Coding: Dominant = Y (barrier), Associated = X (
     goal)

Pattern 5: SINGLE-ISSUE (NONE - use sparingly)
Structure: "[Only DOMAIN_X mentioned, no cross-
     domain link]"
Logic: Genuinely isolated concern
Coding: Dominant = X, Associated = NONE

Note: Patterns 1-4 are COMMON. Pattern 5 (NONE)
     is EXCEPTION.
```

A.3 Strategy C: Two-Step Listing

Strategy C instructs the model to enumerate all relevant domains before prioritizing:

```
STEP 1: ENUMERATE ALL RELEVANT DOMAINS
Before deciding on dominant/associated, first
     list ALL
NASSS domains that the feedback touches:
- Read the entire feedback carefully
- For EACH of the 7 domains, ask: "Does this
     feedback
  mention or imply anything related to this
     domain?"
- Create a mental list of all relevant domains

STEP 2: PRIORITIZE AND ASSIGN
From your enumerated list:
- Dominant Domain: Which captures the PRIMARY
     issue?
- Associated Domain:
  - If list has 2+ domains -> second most
     relevant
  - If list has only 1 domain -> "NONE" (should
     be rare)

SELF-CHECK QUESTIONS BEFORE ASSIGNING NONE:
1. Any IMPACT on workflow/workload? -> Consider
     ADOPTER/ORG
2. Any CAUSE or REASON mentioned? -> May be
     different domain
3. Any WHO is affected? -> Consider ADOPTER
4. Any EXTERNAL constraints? -> Consider
     WIDER_SYSTEM
5. Any PATIENT-related factors? -> Consider
     CONDITION

If YES to any above, you likely need an
     Associated Domain.
```

A.4 Strategy D: Enum-First Output Format

Strategy D modifies the JSON output schema to require enumeration before selection:

```
CRITICAL: List ALL relevant domains BEFORE
     selecting.

The output format requires:
1. First, populate all_relevant_domains with
     EVERY domain
   this feedback touches (1-4 domains)
2. Then, select dominant_domain from that list

Standard Output Schema:
{
  "dominant_domain": "2_TECHNOLOGY",
  "associated_domain": "NONE"  // Easy to
    default here
}

Enum-First Output Schema:
{
  "all_relevant_domains": ["2_TECHNOLOGY", "5
    _ORGANIZATION"],
  "dominant_domain": "2_TECHNOLOGY"
  // Associated derived: list[1] if len >= 2,
    else "NONE"
}

Key Principle:
- The list forces consideration of ALL
    possibilities
- If 2+ domains listed, second becomes "
    associated"
- Model cannot skip enumeration step

Scoring Guide for Inclusion:
- Include domain if relevance >= 2/10
- Exclude only if relevance is truly 0-1/10
```

The key insight is that requiring a list *before* selection structurally prevents the model from defaulting to single-label classification.

B Few-Shot Sensitivity Analysis

We conducted preliminary experiments with few-shot prompting using Magistral to validate the example bias concern discussed in the main paper. Three configurations were tested: *minimal* (few examples), *guided* (curated examples), and *full* (comprehensive examples). To examine fine-grained bias effects, we further distinguished subdomains within Technology: UI/features (2A) and AI/accuracy (2B).

Metric	Minimal	Guided	Full
Technology %	90.9%	85.8%	84.0%
NONE Rate	78.6%	73.5%	73.2%
Subdomain Distribution (within Technology):			
2A (UI/features)	28.1%	46.6%	52.9%
2B (AI/accuracy)	55.0%	45.4%	43.6%

Table 7: Few-shot sensitivity analysis. While few-shot reduces NONE rate, subdomain distribution shifts substantially based on example composition (2A: +24.8%p from minimal to full).

These results confirm the example bias trade-off: few-shot prompting improves multi-label detection (NONE rate: 78.6% → 73.2%) but introduces substantial bias in subdomain classification. The zero-shot Enum-First approach avoids this bias while achieving better multi-label rates (NONE rate: 19.1%).