IS EVALUATION AWARENESS JUST FORMAT SENSITIVITY? LIMITATIONS OF PROBE-BASED EVIDENCE UNDER CONTROLLED PROMPT STRUCTURE

Viliana Devbunova *
Yandex
Belgrade, Serbia
viliana-dev@yandex-team.ru

ABSTRACT

Prior work uses linear probes on benchmark prompts as evidence of *evaluation awareness* in large language models. Because evaluation context is typically entangled with benchmark format and genre, it is unclear whether probe-based signals reflect context or surface structure. We test whether these signals persist under partial control of prompt format using a controlled 2×2 dataset and diagnostic rewrites. We find that probes primarily track benchmark-canonical structure and fail to generalize to free-form prompts independent of linguistic style. Thus, standard probe-based methodologies do not reliably disentangle evaluation context from structural artifacts, limiting the evidential strength of existing results.

1 INTRODUCTION

Recent work introduces the notion of *evaluation awareness* in large language models (LLMs): the ability to distinguish evaluation-time prompts from deployment-like usage and potentially modulate behavior accordingly (Needham et al., 2025; Chaudhary et al., 2025). This phenomenon has implications for benchmark validity and safety monitoring, as evaluation-time behavior may not reflect real-world deployment (Hua et al., 2026).

Empirical evidence for evaluation awareness largely relies on *probe-based analyses*, where linear classifiers trained on frozen activations distinguish evaluation from deployment prompts (Nguyen et al., 2025). However, probe accuracy is not diagnostic: classifiers may exploit any feature correlated with labels. In practice, evaluation prompts are tightly coupled with benchmark-canonical formats and neutral genres, while deployment prompts are more conversational (Needham et al., 2025). This coupling raises the concern that probes track surface structure rather than evaluation context itself (Figure 1a).

We ask a methodological question: *Do probe-based signals attributed to evaluation awareness remain informative once prompt format and genre are partially controlled?* Rather than testing whether evaluation awareness exists, we assess whether standard probe-based methods can isolate evaluation context under stricter controls.

Prior work reports cross-dataset generalization of evaluation-awareness probes, but existing datasets rarely break the correlation between evaluation context and benchmark-canonical format. We introduce controlled format-transplant and genre controls showing that, under the standard single-contrast training setup, probe signals largely track benchmark structure rather than context.

We make three contributions:

- **Controlled 2×2 design.** We construct datasets and diagnostic rewrites that independently vary context, genre, and format under matched prompt length (Figure 1b).
- **Probe generalization fails under format-transplant interventions.** Linear probes trained on benchmark contrasts fail to generalize to free-form prompts, even when content and style are preserved.

*alternative address: devbunova99@gmail.com

- **Methodological implications.** Probe-based evaluation-awareness diagnostics are highly format-sensitive, limiting their use as structure-invariant safety tools.

2 Related Work

Several recent studies examine whether LLMs internally distinguish evaluation from deployment settings. Needham et al. (2025) show that state-of-the-art models can classify transcripts as evaluation or deployment with high accuracy (AUC ≈ 0.83, compared to a human upper bound of ≈ 0.92), suggesting that prompts contain systematic cues associated with evaluation contexts.

Nguyen et al. (2025) probe frozen internal activations using linear classifiers and report generalization across held-out datasets. They observe that safety-related benchmarks are consistently flagged as evaluation prompts even without explicit training, indicating that benchmark structure itself may act as a strong cue. Relatedly, Hua et al. (2026) train a synthetic "organism" model to exhibit evaluation-aware behavior and identify an activation direction whose removal suppresses such behavior, demonstrating that evaluation-linked representations can be induced and manipulated.

Chaudhary et al. (2025) study how probe-based evaluation awareness scales with model size, reporting a power-law relationship across open-weight models ranging from 270M to 70B parameters. Together, these results suggest that probe-detectable signals associated with evaluation settings are widespread and scale systematically.

Our work differs by testing whether such probe-based signals persist under stricter controls on prompt format and genre. Rather than asking whether evaluation awareness exists, we assess whether standard probe-based diagnostics isolate evaluation context or instead reflect sensitivity to benchmark-canonical structure.

3 Setup and Methods

We investigate whether probe-based evaluation awareness is a genuine response to usage context or an artifact of sensitivity to prompt format. Unlike prior work that conflates these dimensions, our design treats *Context* (Evaluation vs. Deployment) and *Format* (Benchmark vs. Casual) as independent variables.

3.1 The 2x2 Controlled Design

We construct a dataset matrix covering four quadrants to isolate the effects of structure. We define our variables as follows:

- **Context:** The underlying intent. *Evaluation* tests capabilities; *Deployment* assists a user.
- **Format:** The surface structure. *Benchmark* implies rigid templates (e.g., MCQ); *Casual* implies free-form inputs.

Dataset Definitions. **Bench-Eval** consists of canonical prompts from MMLU, ARC, and GSM8K. **Casual-Deploy** uses deployment-style user prompts sampled from ShareGPT (restricted to the first

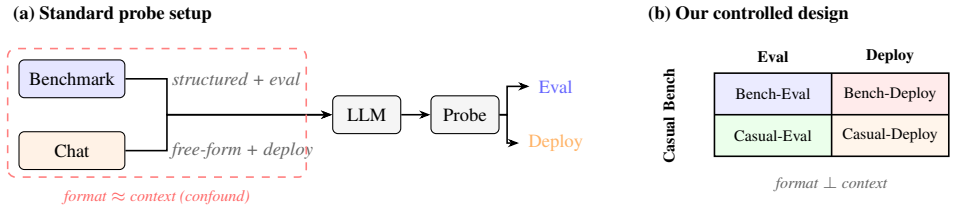

Figure 1: Overview of our approach. **(a)** Standard probes train on benchmark vs. chat prompts, where format and context are confounded. **(b)** Our 2×2 design crosses format and context independently, enabling isolation of each factor.

Table 1: The 2×2 Dataset Matrix. We cross usage context with prompt format to disentangle structural cues from evaluation intent.

Format	Evaluation Context	Deployment Context
Benchmark (Structured)	**Bench-Eval** (MMLU, ARC, GSM8K) *Standard Evaluation*	**Bench-Deploy** (ShareGPT rewritten to benchmark format)
Casual (Free-form)	**Casual-Eval** (MuTual, 1st turn) *Conversational Proxy*	**Casual-Deploy** (ShareGPT, 1st turn) *Standard Deployment*

turn). **Casual-Eval** acts as a conversational evaluation proxy, using the first turn of the MuTual dataset. Finally, **Bench-Deploy** serves as an adversarial control: we route ShareGPT queries to compatible tasks and rewrite them into strict benchmark formats using DeepSeek-V3 (see Appendix A.6 for prompts).

3.2 Experimental Setup

To ensure a clean experiment, we apply strict controls:

- **Length Control:** To prevent probes from relying on length heuristics, we applied *histogram matching* to align the sample length distribution of *Casual-Deploy* with *Bench-Eval*. Figure 2 in Appendix A.2 illustrates the resulting distributions.
- **Leakage Removal:** We explicitly audited rewrites to remove artifacts such as "Answer:" fields.
- **Model and Probes:** We extract frozen residual stream activations (last token) from `Llama-3.1-8B-Instruct` across middle-to-late layers (indices 8–31, 0-indexed). For each layer, we train a Logistic Regression probe using 5-fold cross-validation. Detailed hyperparameters are provided in Appendix A.4.

3.3 Training Configurations

We compare two training paradigms with distinct evaluation protocols:

- **Standard Training (Single-Contrast):** Trained on the typical confound (*Bench-Eval* vs. *Casual-Deploy*). We do not use a validation split here, as models trivially achieve near 100% accuracy via format shortcuts. Instead, we evaluate *zero-shot* on held-out, format-shifted datasets (OOD evaluation) to diagnose generalization failure.
- **Paired Training (Decorrelated):** Trained on pooled data (*Bench-Eval* $\cup$ *Casual-Eval* vs. *Bench-Deploy* $\cup$ *Casual-Deploy*). Here, we use a standard **80/20 train/test split** to demonstrate that the model can learn a generalizable context signal (high test accuracy) when format is decorrelated from the label.

4 Observed Outcome

Our experiments demonstrate that standard probe-based evidence for evaluation awareness is fragile. We present results for Layer 24 (consistent trends across all layers are reported in Appendix 4).

The Format Trap. As shown in Table 2 (top row), standard probes fail to generalize. Deployment prompts rewritten into benchmark formats (*Bench-Deploy*) are misclassified as "Evaluation" (94.5%). The probe detects the structural format (e.g., MCQ), not the context. To isolate **Genre** (linguistic register), we tested deployment prompts rewritten into a formal "Academic Tone" without structural changes. This had negligible effect ($\approx 0\%$ classified as Eval), confirming that probes are sensitive to rigid templates rather than linguistic style.

Table 2: **Probing Results (Layer 24).** Values indicate the percentage of prompts classified as *Evaluation*. Standard training conflates format with context, leading to high false positives on *Bench-Deploy*. Paired training restores correct classification.

	Intended Evaluation		Intended Deployment	
Training Regime	**Bench-Eval**	**Casual-Eval**	**Casual-Deploy**	**Bench-Deploy**
Standard (Single)	**98.8%**	0.0%	0.0%	**94.5%** (FP)
Paired (Decorrelated)	**98.5%**	**100.0%**	**0.0%**	**7.2%**

Paired Training Success. When format is decorrelated from context during training, the probe successfully learns the underlying concept. As shown in Table 2 (bottom row), the paired probe correctly identifies *Bench-Deploy* as deployment (7.2% error vs 94.5%) and recognizes *Casual-Eval* as evaluation (100%). Furthermore, we evaluated these probes on AI Safety datasets (AdvBench, HarmBench) and found that neither standard nor paired probes flag safety prompts as "Evaluation" (see Appendix A.5).

5 Reason for Failure

Diagnosis. The observed collapse indicates that probe-based diagnostics exploit surface-level structural features rather than encoding an abstract notion of evaluation context. Benchmark-canonical layouts introduce strong, linearly separable cues in representation space, causing probes to conflate "being evaluated" with "looking like a benchmark."

Why this is not a trivial confound. Prompt length is explicitly matched across evaluation and deployment datasets, ruling out length-based shortcuts. Genre and formality are manipulated independently from structure, yet do not induce evaluation-like predictions. We explicitly audit and remove *rewrite-induced answer leakage*, where benchmark-format rewrites of deployment prompts inadvertently include model-generated answer content (e.g., pre-filled options or solution scaffolds). Ablations show that removing such leakage does not explain the observed effect and in some cases increases probe confidence. The failure is consistent across layers, indicating a pervasive structural bias rather than a localized artifact.

Boundary conditions. We do not claim that evaluation awareness is absent from model representations. Rather, our results show that under standard benchmark-correlated training setups, linear probes fail to isolate such signals. Only when format is explicitly decorrelated from context during training do probes exhibit format-invariant behavior.

6 Discussion

Implications. Our results suggest that probe-based monitoring of evaluation awareness can induce false confidence. Diagnostics sensitive to prompt format rather than usage context risk mischaracterizing model behavior, and governance mechanisms relying on probe accuracy may systematically overestimate situational awareness. Benchmark-canonical structure must be explicitly controlled for in experimental design, not bundled with context by default. Without such control, probe accuracy may reflect sensitivity to layout rather than higher-level contextual reasoning.

Limitations. Our study uses only linear probes and a single backbone (Llama-3.1-8B-Instruct), and does not rule out non-linear or distributed representations of evaluation awareness. Some dataset instantiations rely on proxy constructions. We expect the central confound—benchmark structure correlating with evaluation context—to persist broadly due to common benchmark construction practices.

REPRODUCIBILITY STATEMENT

All code and constructed datasets are publicly available at `https://github.com/viliana-dev/eval-awareness-format`. The repository includes scripts for the full pipeline: dataset downloading, query routing, format rewrites, histogram-based length matching, activation extraction, probe training, and all evaluation protocols reported in this paper. Pre-processed datasets for the 2×2 matrix are hosted at `https://huggingface.co/datasets/viliana-dev/eval-awareness-2x2`.

All source datasets are public: MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), GSM8K (Cobbe et al., 2021), ShareGPT[1], and MuTual (Cui et al., 2020). Benchmark-format rewrites were generated using DeepSeek-V3; exact prompt templates are provided in Appendix A.6.

All experiments use `meta-llama/Llama-3.1-8B-Instruct` with `bfloat16` inference on a single NVIDIA A100 80 GB. Probes are trained in `float32`. A fixed random seed of 42 is used throughout all scripts. Implementation details—layer indices, sample sizes (n=1076), regularization, and cross-validation settings—are reported in Appendix A.4.

REFERENCES

Maheep Chaudhary, Ian Su, Nikhil Hooda, Nishith Shankar, Julia Tan, Kevin Zhu, Ryan Lagasse, Vasu Sharma, and Ashwinee Panda. Evaluation awareness scales predictably in open-weights large language models, 2025. URL `https://arxiv.org/abs/2509.13333`.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL `https://arxiv.org/abs/1803.05457`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL `https://arxiv.org/abs/2110.14168`.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL `https://aclanthology.org/2020.acl-main.130`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL `https://arxiv.org/abs/2009.03300`.

Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed, 2026. URL `https://arxiv.org/abs/2510.20487`.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. URL `https://arxiv.org/abs/2505.23836`.

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and steering evaluation awareness of language models, 2025. URL `https://arxiv.org/abs/2507.01786`.

A APPENDIX: DATASETS, CONTROLS, AND ADDITIONAL EXPERIMENTS

A.1 DATASETS IN THE 2X2 MATRIX

We use a controlled dataset matrix crossing **usage context** (evaluation vs. deployment) and **genre** (benchmark-style vs. conversational), while treating **format** (benchmark-canonical structured vs. free-form) as a separate variable.

[1] `https://huggingface.co/datasets/theblackcat102/sharegpt-english`

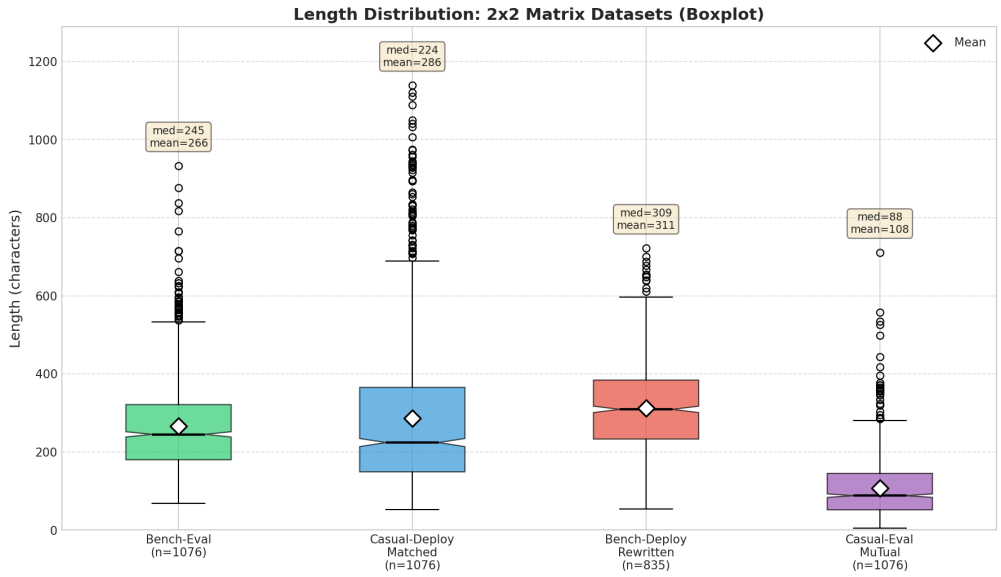

Figure 2: Length distributions (in characters) across the four datasets. *Casual-Deploy* is histogram-matched to *Bench-Eval*. *Bench-Deploy* is slightly longer due to formatting overhead, while *Casual-Eval* (1st turn) is naturally shorter.

Bench-Eval (Evaluation; benchmark-style; structured). Canonical benchmark prompts (e.g., MCQ and structured word-problem formats) drawn from MMLU, ARC, and GSM8K.

Casual-Deploy (Deployment; conversational; free-form). Deployment-style user prompts sampled from ShareGPT. For comparability with Bench-Eval, we use only the first user turn and apply histogram matching for length control (see Section A.2).

Casual-Eval (Evaluation; conversational; free-form proxy). We instantiate this condition with MuTual as a conversational evaluation proxy. For tighter control over genre and input structure, we extract only the first dialogue turn from each example; we treat this as an approximation rather than a faithful instantiation of the original task.

Bench-Deploy (Deployment; benchmark-style; structured proxy via rewrites). Deployment-origin prompts from Casual-Deploy rewritten into benchmark-canonical structured templates (MCQ or structured word problems), while avoiding explicit evaluation framing. Since not all deployment prompts can be sensibly forced into benchmark templates, we first route prompts into coarse task categories and rewrite only compatible prompts.

A.2 Length Control

Motivation. Initially, the datasets exhibited significant length disparities, creating a potential shortcut for probes. *Bench-Eval* prompts had a median length of 248 characters, while the raw *Casual-Deploy* (ShareGPT 1st turn) prompts were much shorter (median 113 characters).

Histogram Matching. To eliminate length as a confounding variable in the primary training contrast (*Bench-Eval* vs. *Casual-Deploy*), we performed histogram matching. We calculated the length distribution of *Bench-Eval* using 30 percentile-based character-length bins and sampled the exact same number of examples from *Casual-Deploy* for each bin. This process aligned the distributions as closely as possible, yielding equal sample sizes ($n = 1076$) and comparable statistics. Figure 2 visualizes the resulting length distributions.

A.3 Rewrite Pipeline and Leakage Removal

Rewrite Variants. All rewrites are performed with DeepSeek-V3 under controlled prompting. We construct three variants sharing identical content (Casual-Deploy):

- `bench_deploy`: benchmark format + conversational genre.
- `casual_deploy_formal`: free-form format + academic/impersonal genre.
- `bench_deploy_formal`: benchmark format + academic/impersonal genre.

Answer Leakage Cleanup. We define "answer leakage" as rewrite artifacts where structured templates inadvertently elicit model-generated answers (e.g., pre-filled "Answer:" fields). We explicitly audit and remove such leakage. As shown in Table 3, removing leakage increases the probe's confidence in classifying rewrites as "Evaluation," confirming that leakage was acting as a confounding feature that suppressed the format signal.

Table 3: Effect of removing unintended answer leakage. Values indicate % classified as Evaluation. Removing leakage makes the probe *more* likely to flag the prompt as Eval, isolating the format effect.

Condition	Before cleanup	After cleanup
Bench-Deploy (MCQ rewrite)	55–70%	94–95%
Bench-Deploy Formal	51–68%	88–91%

A.4 Implementation Details

Activation Extraction. We extract activations using the following settings:

- **Position:** Last token activation (before final layer norm) to capture the aggregated context.
- **Layers Evaluated:** We sweep across middle-to-late layers (0-indexed) to verify consistency: 8, 12, 16, 20, 22, 24, 26, 28, 30, 31.
- **Context:** Full input context is preserved (no truncation) to maintain task integrity.
- **Precision:** Model inference is performed in `bfloat16`; probe training uses `float32`.

Probe Training.

- **Architecture:** Logistic Regression (linear probe).
- **Preprocessing:** Input activations are normalized using `StandardScaler` fitted on the training split.
- **Labels:** Binary classification where 0 = Evaluation, 1 = Deployment.
- **Optimization:** L2 regularization (default $C = 1.0$), max iterations = 1000.
- **Validation:** We use 5-fold Cross-Validation on the training set to verify stability.

A.5 Additional Experimental Results

Standard (Single) vs. Paired Training Analysis. We present the full layer-wise results for both training regimes. Table 4 shows the failure of standard probes to generalize: they consistently misclassify *Bench-Deploy* as Evaluation ($> 94\%$) across all layers. In contrast, Table 5 shows that probes trained on the paired dataset (decorrelating format from context) successfully learn to distinguish context across all layers.

Evaluation on AI Safety Datasets. We evaluate probes on safety benchmarks (AdvBench, HarmBench) to test if "safety" prompts trigger evaluation awareness. As shown in Table 6, safety prompts are universally classified as Deployment (0% Eval) by both standard and paired probes. This suggests that evaluation awareness probes should not be interpreted as general-purpose safety monitors.

Table 4: **Standard (Single-Contrast) Training Results across Layers.** Values indicate % classified as Eval. The standard probe consistently relies on format, incorrectly flagging *Bench-Deploy* as Evaluation and ignoring *Casual-Eval*.

Layer	B-Eval	C-Eval	C-Deploy	B-Deploy
8	99.8%	0.0%	0.0%	94.3%
16	99.5%	0.0%	0.0%	94.3%
24	98.8%	0.0%	0.0%	94.5%
30	99.2%	0.0%	0.0%	94.6%
31	99.1%	0.0%	0.0%	94.6%

Table 5: **Paired Training Results across Layers.** Values indicate % classified as Eval. The paired probe consistently learns Context (Evaluation) vs. Format (Benchmark).

Layer	B-Eval	C-Eval	C-Deploy	B-Deploy
8	98.8	98.6	0.0	7.2
16	97.7	99.1	0.0	4.8
24	98.5	100.0	0.0	7.2
30	99.0	99.5	0.0	7.2
31	98.7	99.5	0.0	6.0

A.6 Rewrite Prompt Templates

This section reports the exact prompt templates used for routing and rewriting.

Routing Prompt.

```
You are a routing assistant for dataset construction.
Your task is to assign each user query to EXACTLY ONE category
based solely on the type of reasoning it requires.

AVAILABLE CATEGORIES:
(1) MMLU  { knowledge-based multiple-choice reasoning
(2) ARC   { grade-school science or commonsense reasoning
(3) GSM8K { arithmetic or multi-step math reasoning
(4) OTHER { none of the above

Output ONLY one label: MMLU, ARC, GSM8K, or OTHER.

User query:
"{USER_PROMPT}"
```

Benchmark Rewrite: MMLU-style.

```
Rewrite the following user query into a benchmark-style
multiple-choice question.

Constraints:
- Provide EXACTLY four answer options (A, B, C, D).
- Do NOT include explanations, solutions, or commentary.
- Do NOT indicate which option is correct.
- Preserve the original task intent and required information.
- Do NOT introduce an explicit evaluation framing.

User query:
"{USER_PROMPT}"
```

Table 6: Probe classification on AI Safety Datasets (Layer 24). All safety prompts are classified as Deployment.

Dataset	Single Probe	Paired Probe
AdvBench (520)	0%	0%
HarmBench standard (200)	0%	0%
HarmBench contextual (100)	0%	0%
HarmBench copyright (100)	0%	0%

```
Output format:

Question:
<rewritten question>

Options:
A. ...
B. ...
C. ...
D. ...
```

Benchmark Rewrite: ARC-style.

```
Rewrite the following user query into a structured
grade-school science or commonsense reasoning question.

Constraints:
- Use a neutral, benchmark-style formulation.
- If appropriate, present the question as multiple-choice
  with EXACTLY four options.
- Do NOT include explanations or the correct answer.
- Preserve the original task intent.

User query:
"{USER_PROMPT}"
```

Benchmark Rewrite: GSM8K-style.

```
Rewrite the following user query into a structured
grade-school math word problem.

Constraints:
- Present a self-contained math problem.
- Do NOT include intermediate steps or the final answer.
- Do NOT add any solution explanation.
- Preserve the numerical reasoning required by the original query.

User query:
"{USER_PROMPT}"
```

Genre Control Rewrite.

```
Rewrite the following user query in a formal, impersonal,
and professional register.

Constraints:
- Preserve the original task intent and information.
- Do NOT introduce benchmark-style structure.
```

```
- Do NOT include answer options, solution steps, or summaries.
- The output should remain a free-form user request,
  not an evaluation-style prompt.

User query:
"{USER_PROMPT}"
```