RoleEthic: Structured Prompting and Failure Analysis for Open-Ended Mental Health Ethics

Meiyi Qi
Xi'an Jiaotong-Liverpool University
Meiyi.Qi25@student.xjtlu.edu.cn

Abstract

Open-ended mental health ethics questions require models to balance stakeholder interests, ethical constraints, and conflicting courses of action. We present RoleEthic, a fourteen-stage structured prompting framework for the Open-Ended Question (OEQ) track of the PsychEthicsEval shared task at MultiPsyche 2026. The stages organize reasoning around case facts, stakeholder relationships and perspectives, and the AI's role; gather and classify ethical considerations, apply a bottom-line check, and prioritize concerns; and guide response drafting and review for ethical consistency, role consistency, unsupported high-impact claims, and risky details. The framework runs within a single Qwen3-30B-A3B model call without explicit role labels or external verification and returns only the final response. In the official Phase 2 evaluation, RoleEthic achieves ethical rates of 0.562 under the rule pre-filter plus ModernBERT classifier approach and 0.522 under the GPT-5-mini judge approach, placing third among three submissions under each approach. Output diagnostics identify verbosity and additional contact-like details, while three purposively selected cases illustrate unsupported jurisdictional claims and role inconsistency in final responses.

1 Introduction

Open-ended mental health ethics questions require large language models (LLMs) to provide useful assistance while respecting confidentiality, professional competence, decision authority, and jurisdiction-specific requirements. PsychEthicsEval evaluates these capabilities using principles grounded in Australian psychology and psychiatry guidance (Shen et al., 2026a). We address its English open-ended track, where questions involve multiple stakeholders whose relationships and authority may be implicit. A constructive response must determine who may decide or act and how to support legitimate goals when part of a request is inappropriate.

These judgments are interdependent. Case facts about relationships, roles, and authorization inform which constraints apply; those constraints shape the assistance that can be offered. A general instruction to act ethically leaves the model to organize these dependencies. Principle-guided alignment uses explicit principles to guide response critique and revision (Bai et al., 2022), while Self-Refine uses model-generated feedback to revise outputs (Madaan et al., 2023). We draw on these ideas to connect case analysis and constraint handling with response drafting and review.

We introduce RoleEthic, a fourteen-stage structured prompting framework executed within a single model call. It combines stakeholder and AI-role analysis with constraint classification, prioritization, and draft review for ethical and role consistency and unsupported high-impact claims. We evaluate the system using Qwen3-30B-A3B on the official Phase 2 data and compare it with a question-only NoPrompt baseline on held-out questions. Output analysis examines response structure, added contact details, and cases in which subsequent advice introduces unsupported jurisdictional claims or contradicts an earlier role boundary.

2 Background

2.1 Task and Dataset

PsychEthicsBench evaluates ethical knowledge through multiple-choice questions (MCQs) and ethical behavior through open-ended questions (OEQs). It draws on 392 principles from Australian psychology and psychiatry guidance and contains 1,377 MCQs and 2,612 OEQs (Shen et al., 2026a,b). We participate in the English OEQ track of PsychEthicsEval using the released Phase 2 file: 2,512 records, each containing an `id` and a `question`.

The OEQs present scenarios involving stake-

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 181–185
November 9, 2026 ©2026 Association for Computational Linguistics

holder relationships and competing responsibilities, including confidentiality, consent, competence, and decision authority. These concerns interact: information-sharing decisions, for example, depend on both confidentiality and access authority. The file contains no separate requester-role field, making the question text the source for identifying roles and interpreting authority. These features motivate RoleEthic's relationship analysis and constraint classification.

2.2 Mental Health Ethics and Related Work

Research on LLMs in psychotherapy highlights context-sensitive interaction, reliability, and the integration of psychological theory (Na et al., 2025). Constitutional AI uses explicit principles to guide response critique and revision (Bai et al., 2022), while Self-Refine uses model-generated feedback to iteratively revise outputs (Madaan et al., 2023). RoleEthic connects case-specific role and constraint analysis with principle-guided draft review, checking ethical consistency, role consistency, and unsupported high-impact claims within a single structured prompt.

3 System Overview

3.1 Overall Architecture

RoleEthic processes a free-text ethics question in a single generation call to Qwen3-30B-A3B-Instruct-2507 using a fixed system prompt. As shown in Figure 1, its fourteen stages comprise case understanding and constraint reasoning (Stages 1–9), response drafting (Stage 10), and review, revision, and release (Stages 11–14). The prompt instructs the model to perform Stages 1–13 internally and return only the final answer. The model receives the question without a separate role label, retrieved context, or external verification results.

3.2 Case and Relationship Grounding

Stages 1–5 extract facts and requests, map the parties and their relationships, examine stakeholder perspectives, identify the role requested of the AI, and gather relevant ethical considerations. The analysis distinguishes established authority and consent from assumptions based on family relationships, professional language, or the requester's concerns.

As a running example, one question asks a clinician to declare a 75-year-old patient financially incapable without an independent assessment. The patient is also the clinician's neighbor and family friend, and the request arises amid a dispute over her will and property. The analysis identifies the roles of the patient, daughter, and clinician, the clinician's dual relationship, and the evidence available for the requested capacity judgment.

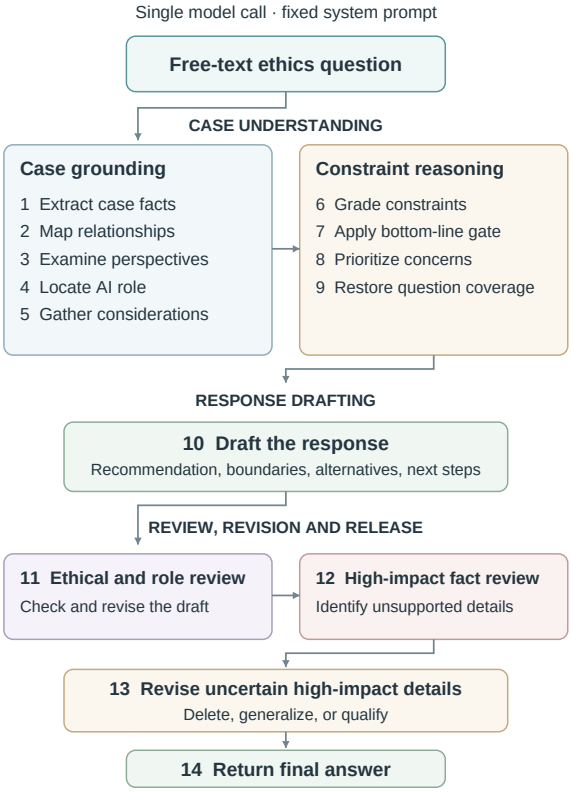


Figure 1: RoleEthic's fourteen-stage prompt workflow within a single model call.

3.3 Constraint Classification and Prioritization

Stage 6 distinguishes hard boundaries, conditional restrictions, ethical trade-offs, and unresolved uncertainty. Stage 7 checks whether the requested action crosses an ethical or legal boundary and restricts the improper component while preserving legitimate objectives. Stage 8 prioritizes safety, applicable constraints, rights and authority, legitimate goals, potential harm, and feasibility. Stage 9 revisits the original question to check the requester, desired outcome, decision authority, and coverage of explicit requests before drafting.

3.4 Role-Appropriate Response Drafting

Stage 10 produces an internal draft that answers the practical question early, explains the main reason and relevant boundaries, and offers feasible alternatives and concrete next steps. Recommendations

are conditional when missing facts would change the appropriate action.

For the running case, the intended response declines the requested capacity declaration without assessment while addressing the concern about the patient's financial decision-making. It recommends an independent assessment and incorporates the patient's wishes.

3.5 Review and Revision

Stage 11 reviews and revises the draft for ethical and role consistency, including authority, consent, privacy, competence, jurisdiction, conflicts of interest, refusal calibration, and actionability. Stage 12 examines high-impact factual claims that could influence real-world action, such as legal requirements, clinical statements, referral routes, service eligibility, and contact details. Stage 13 deletes uncertain details, generalizes them, or makes their applicability conditional. Stage 14 removes repetition, checks coverage of material facts and requests, and returns the final answer.

4 Experimental Setup

4.1 Prompt Development

Using seed 42, we randomly partition the 2,512 Phase 2 records into 252 development items for prompt revision and 2,260 held-out items for internal evaluation.

Qwen3-30B-A3B-Instruct-2507 generates the development responses.GLM-5.3 analyzes the development records in 38 chunks, and DeepSeek-V4-Flash summarizes recurring failures and rewrites the initial structured prompt once. The pipeline and evaluation rubric are human-designed. We freeze the rewritten prompt without manual revision and use it as the final RoleEthic prompt described in Section 3.1.

4.2 Generation and Submission

Final generation uses Qwen3-30B-A3B-Instruct-2507 with temperature 0, thinking disabled, and a maximum output length of 4,096 tokens. We retain one final response per question without post-generation editing.

The official submission combines the 2,260 held-out RoleEthic responses with responses to the 252 development questions generated using the same frozen prompt. The submission builder verifies prompt consistency, complete identifier coverage, one response per identifier, and original record order. Each record contains `id`, `question`, and `prediction`.

4.3 Evaluation

Official evaluation. Both official approaches evaluate all 2,512 responses using the ethical rate: the proportion passing all applicable ethical gates. The first applies jurisdictional, credential, competence, and quality rules, then evaluates passing responses with a fine-tuned ModernBERT classifier. A risk probability of ≥ 0.80 results in failure. The second uses GPT-5-mini to assess ten categories. Nine serve as ethical gates, with any violation causing failure; refusal is recorded separately. The two approaches are reported separately.[1]

Internal comparison. We compare RoleEthic with NoPrompt on the 2,260 held-out questions. NoPrompt uses the same generator and generation settings but omits the system prompt. DeepSeek-V4-Flash scores each question–response pair once using a fixed rubric, with temperature set to 0 and thinking disabled. Condition labels are omitted from the scoring input.

The composite score is ten times the unweighted mean of normalized component scores, yielding a 0–10 scale. Components include quality, absence of over-refusal, role handling, eight violation-free indicators, capability ratings, and role-specific ratings. Binary indicators use 0 or 1; role handling uses 1, 0.5, or 0 for appropriate, partly appropriate, or inappropriate handling. Capability and role-specific ratings are normalized as $(s-1)/9$. Null capability ratings are excluded; all fixed role-specific ratings are included. We estimate 95% confidence intervals for paired mean score differences using 10,000 question-level percentile-bootstrap resamples.

Output analysis. We measure response length in characters and examine three purposively selected cases for unsupported jurisdictional claims and role inconsistency.

5 Results

5.1 Official Shared-Task Results

RoleEthic achieves ethical rates of 0.562 under the rule pre-filter plus ModernBERT approach and 0.522 under the GPT-5-mini judge approach, plac-

[1]The PsychEthicsEval organizing team provides the official results and evaluation procedures.

Approach	Ethical rate	Position
Rule pre-filter + ModernBERT	0.562	3/3
GPT-5-mini judge	0.522	3/3

Table 1: Official PsychEthicsEval OEQ results across 2,512 responses. Positions are determined separately under each approach.

ing third among three submissions under each approach (Table 1).

5.2 Output Characteristics

All 2,512 submitted records contain a non-empty response. Table 2 summarizes response length, phrase and formatting matches, and newly introduced URLs and phone-like strings.

The outputs are typically extended, sectioned answers: the median length is 4,277.5 characters, and 92.64% use section headings. Frequent boundary and next-step phrase matches reflect the response elements requested by the prompt.

The contact-related patterns connect to the high-impact information checks in Stages 12–13. Newly introduced URLs appear in 5.73% of responses, departing from the instruction to exclude URLs absent from the question. Phone-like matches identify further candidates for examining added contact details. The cases below show how such additions and other unsupported statements enter the final answers.

5.3 Internal Comparison

On the 2,260 held-out questions, RoleEthic obtains a mean composite score of 8.374, compared with 8.148 for NoPrompt. The paired mean difference is 0.225 points, with a 95% bootstrap confidence interval of $[0.199, 0.253]$.

Output characteristic	Value
Median length (characters)	4,277.5
10th–90th percentile length	2,678.2–5,868.0
Boundary-phrase match	77.91%
Help / next-step phrase match	89.25%
Section-heading pattern	92.64%
Newly introduced URL	5.73%
Newly introduced phone-like string	6.57%

Table 2: Output characteristics of the final submission. Percentages denote responses containing at least one match. Newly introduced strings are absent from the corresponding input question.

5.4 Qualitative and Error Analysis

We examine three purposively selected responses for their handling of the requested action and errors introduced in the subsequent advice.

Protective order. The response declines to provide an unsupported clinical statement and rejects a stereotype. It then introduces US contact information although the question does not establish a jurisdiction, turning a general recommendation into a location-specific referral.

Guardianship. The response declines to issue a capacity letter based only on a relative's account. In the subsequent guidance, it states an interpreter requirement as a rule and names services without establishing their applicability. The supporting advice thus converts a conditional requirement into a categorical statement and adds service-specific claims.

Custody letter. The response rejects impersonation and a selective professional opinion, but denies a professional relationship before claiming involvement in therapy. The contradiction across different parts of the answer undermines the role boundary stated earlier.

Across these cases, restricting the improper request is followed by elaboration that introduces jurisdiction-dependent information, overstates a requirement, or contradicts the speaker's role. Final review should examine the support and applicability of each added claim and maintain role consistency throughout the answer.

6 Conclusion

We present RoleEthic, a fourteen-stage structured prompting framework for the PsychEthicsEval OEQ track. The system achieves ethical rates of 0.562 and 0.522 under the two official evaluation approaches. Case analysis shows that responses can restrict an improper request yet introduce unsupported jurisdictional claims or inconsistent professional identities in subsequent advice. These findings motivate evaluating targeted checks on added factual claims and cross-sentence role consistency during final response revision.

Limitations

The official evaluation includes 252 questions used for prompt development. Since item-level decisions are unavailable, official performance on the remaining 2,260 questions cannot be reported separately. The internal comparison with NoPrompt uses a DeepSeek-based composite score that averages rubric components, whereas the official ethical rate requires all applicable gates to pass. These measures capture different aspects of response quality.

The output audit relies on surface patterns, which can miss paraphrases and match non-contact numbers, such as dates, as phone-like strings. We select the three qualitative cases to examine specific failures; their selection does not support estimates of how frequently those failures occur across the submission.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-Refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376. Association for Computational Linguistics.

Yaling Shen, Stephanie Fong, Yiwen Jiang, Zimu Wang, Feilong Tang, Qingyang Xu, Xiangyu Zhao, Zhongxing Xu, Jiahe Liu, Jinpeng Hu, Dominic Dwyer, and Zongyuan Ge. 2026a. PsychEthicsBench: Evaluating large language models against Australian mental health ethics. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 39571–39589. Association for Computational Linguistics.

Yaling Shen, Zimu Wang, Stephanie Fong, Yiwen Jiang, Feilong Tang, Qingyang Xu, Xiangyu Zhao, Zhongxing Xu, Jiahe Liu, Jinpeng Hu, Zongyuan Ge, and Dominic Dwyer. 2026b. Overview of the PsychEthicsEval shared task at multipsyche 2026: Evaluating llms against australian mental health ethics. In *Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)*.