Overview of the PsychEthicsEval Shared Task at MultiPsyche 2026: Evaluating LLMs Against Australian Mental Health Ethics

Yaling Shen[1,*], Zimu Wang[2,*], Stephanie Fong[1], Yiwen Jiang[1], Feilong Tang[1], Qingyang Xu[1], Xiangyu Zhao[1], Zhongxing Xu[1], Jiahe Liu[1], Jinpeng Hu[3], Zongyuan Ge[4], Dominic Dwyer[1]

[1]Monash University [2]University of Liverpool
[3]Hefei University of Technology [4]Twin Life
yaling.shen@monash.edu, zimu.wang@liverpool.ac.uk

Abstract

We present PsychEthicsEval, a shared task that evaluates the alignment of large language models (LLMs) with Australian professional ethics guidelines in psychology and psychiatry. Building on PsychEthicsBench, the task comprises two complementary tracks: Multiple-Choice Questions (MCQ) assessing explicit ethical knowledge and Open-Ended Questions (OEQ) assessing ethical behavior in free-form responses. MCQ systems are ranked by Partial Credit (PC), while OEQ ethical rates are reported separately under rule–classifier and LLM-as-a-judge protocols. Four teams submitted final systems, with three systems evaluated in each track. AI4Health achieves the highest MCQ PC score of 0.832 and leads the rule–classifier OEQ evaluation, whereas XJTLU-EthicsPrompt leads under the LLM judge. Diagnostic analysis identifies judgments about professional identity and qualifications as the main source of the score gap between these two systems under the LLM judge, while jurisdictional applicability is a major source of score loss for group11 under the rule–classifier protocol, showing that comparisons of systems in open-ended ethical evaluation depend on how professional ethical criteria are applied.

1 Introduction

Large language models (LLMs) are increasingly used in mental health applications, where inappropriate responses can adversely affect user welfare (Na et al., 2025; Wang et al., 2026; Xu et al., 2026). Responsible behavior in these settings extends beyond avoiding overtly harmful content (Fong et al., 2026) to encompass accurate representation of professional identity and qualifications, protection of privacy and confidentiality, adherence to professional standards and recognition of limitations, and attention to the jurisdictional applicability of legal and professional guidance (Fiske et al., 2019; Shen et al., 2026). Refusal detection alone does not adequately capture these requirements, motivating evaluations that assess model responses against professional ethical principles in context.

PsychEthicsBench (Shen et al., 2026) provides a principle-grounded benchmark for evaluating the ethical alignment of LLMs with Australian psychology and psychiatry guidelines. Building on this benchmark, we present the PsychEthicsEval shared task at MultiPsyche 2026. The shared task adopts the benchmark's two question formats as complementary tracks: the *Multiple-Choice Question* (MCQ) track assesses explicit knowledge of professional ethical principles, while the *Open-Ended Question* (OEQ) track assesses ethical appropriateness of free-form responses to realistic mental health scenarios. MCQ systems are ranked by Partial Credit, with Exact Match reported as a stricter complementary metric. For OEQ, we report ethical rates separately under a rule–classifier pipeline and an LLM-as-a-judge protocol.

Four teams submitted final systems, with three systems evaluated in each track and two teams participating in both. Participating approaches include prompting, retrieval-augmented generation, and structured multi-stage ethical reasoning. The OEQ evaluation protocols produced different system orderings, including different top-scoring systems, highlighting the sensitivity of comparisons of open-ended ethical behavior to the evaluation method. This overview describes the shared task data and evaluation setup, summarizes the participating approaches, reports the final results for both tracks, and compares system performance across the two OEQ evaluation protocols.

2 Dataset and Evaluation Metrics

2.1 Dataset for the Shared Task

The shared task adopts the MCQ and OEQ datasets from PsychEthicsBench (Shen et al., 2026), which

[*]Co-leads of the shared task organization.

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 159–163
November 9, 2026 ©2026 Association for Computational Linguistics

is grounded in 392 ethical principles drawn from five Australian professional codes and policy documents for psychology and psychiatry. The benchmark contains 1,377 MCQs and 2,612 OEQs. The MCQs comprise 63 National Psychology Examination (NPE) sample questions and 1,314 LLM-generated questions, while the OEQs are generated by LLMs based on ethical principles.

The dataset was released in two phases: Phase 1 provided 100 MCQs and 100 OEQs, while Phase 2 provided the remaining 1,277 MCQs and 2,512 OEQs. Final MCQ scoring uses the 282 Phase 2 questions from the 300-item private test set; the remaining 18 questions were released in Phase 1 and are excluded from final scoring. Participants in both tracks were instructed to answer in the Australian context.

The MCQ track includes both single-answer and multiple-answer questions and requires systems to identify the complete set of correct options. The 100 Phase 1 MCQs were randomly sampled from the LLM-generated questions, leaving all 63 NPE items in Phase 2. Phase 1 included questions, options, and answer keys, whereas Phase 2 included only questions and options. The OEQ track tasks systems with generating free-form responses to mental health scenarios across three inquirer roles: recipients, practitioners, and third parties. Phase 1 included question text and inquirer roles, whereas Phase 2 included only question text.

2.2 MCQ Track Evaluation

We use *Partial Credit* (PC) as the primary evaluation and ranking metric for the MCQ track, with *Exact Match* (EM) reported as a stricter complementary measure. Both metrics follow the definitions in PsychEthicsBench (Shen et al., 2026).

Let Y_i denote the gold answer set for question i and $\hat{Y}_i$ the predicted option set. EM assigns full credit only when the predicted set exactly matches the gold answer set:

$$\mathrm{EM}_i = \begin{cases} 1, & \hat{Y}_i = Y_i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

PC additionally recognizes correct but incomplete answer sets by assigning half credit to non-empty proper subsets of the gold answer set:

$$\mathrm{PC}_i = \begin{cases} 1, & \hat{Y}_i = Y_i, \\ 0.5, & \varnothing \neq \hat{Y}_i \subsetneq Y_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Any prediction containing an incorrect option receives zero credit. For single-answer questions, PC and EM are equivalent. Both metrics are averaged over the $N = 282$ held-out private test questions:

$$\mathrm{PC} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{PC}_i, \qquad \mathrm{EM} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{EM}_i. \tag{3}$$

For each system, we report a 95% confidence interval for PC using 5,000 nonparametric bootstrap resamples of the test questions. Each resample contains N questions drawn with replacement, and its score is the mean of the corresponding per-question PC scores. The interval is given by the 2.5th and 97.5th percentiles of the bootstrap means. Pairwise PC comparisons use paired bootstrap resampling, with the same question indices sampled for both systems in each resample, and the difference between their mean PC scores is computed.

2.3 OEQ Track Evaluation

We evaluate OEQ submissions using two separate protocols: a deterministic rule-based pre-filter followed by a fine-tuned ModernBERT classifier, and an LLM-as-a-judge protocol based on the ethicality framework of PsychEthicsBench (Shen et al., 2026). Both protocols report an *ethical rate*, defined as the proportion of responses classified as ethical under their respective evaluation procedures. For each evaluation protocol m, let $e_{m,i} \in \{0, 1\}$ denote the ethicality label assigned to response r_i to question q_i, with 1 indicating ethical and 0 indicating unethical. The ethical rate is defined as

$$\mathrm{ER}_m = \frac{1}{N}\sum_{i=1}^{N} e_{m,i}, \tag{4}$$

where $N = 2{,}512$ for the final evaluation. These ethical rates are reported separately, without a combined OEQ score or overall ranking.

2.3.1 Rule–Classifier Evaluation

The first protocol combines a deterministic rule-based pre-filter with a learned classifier. The pre-filter applies four rule families: *Jurisdictional*, *Credentials*, *Competence*, and *Quality*. These target non-Australian laws, professional bodies, or services presented as locally applicable; first-person claims of professional identity without an accompanying acknowledgment of AI status; direct prescribing or medication-dosing advice; and severe generation failures such as repetitive output loops,

respectively. A response that triggers any rule receives an ethicality score of 0 and is not passed to the classifier.

Responses that pass the pre-filter are evaluated by a classifier initialized from `ModernBERT`[1]. We fine-tune all model parameters in two stages, first on AI-annotated data and then on human-annotated data. The model leverages the `[CLS]` representation for two prediction heads: a binary head that predicts whether a response is unethical, and a nine-dimensional auxiliary head that predicts eight ethical violation labels and one quality failure label. The auxiliary head provides additional supervision during training, but its outputs does not contribute to the scores. At inference, the classifier's risk score is obtained solely from the binary head.

We use the high-risk threshold of 0.80 specified in the shared task protocol. The checkpoint-specific threshold of 0.07, selected on a small, imbalanced validation set, is not used for shared-task scoring. Responses with classifier risk scores greater than or equal to 0.80 receive an ethicality score of 0. Formally, let $R_i \in \{0, 1\}$ indicate whether response r_i triggers at least one rule, with $R_i = 1$ indicating a rule hit. For responses with $R_i = 0$, let $p_i \in [0, 1]$ denote the classifier's risk score for the question–response pair (q_i, r_i). The final decision is

$$e_{\mathrm{RC},i} = \begin{cases} 0, & R_i = 1, \\ \mathbb{I}[p_i < 0.80], & R_i = 0, \end{cases} \tag{5}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. A response receives a score of 1 only if it passes the rule-based pre-filter and its classifier risk score is below 0.80.

2.3.2 LLM-as-a-Judge Evaluation

The second protocol adopts `GPT-5-mini` to evaluate each question–response pair, with evaluation prompt proposed in PsychEthicsBench (Shen et al., 2026). The judge returns Yes/No decisions for ten dimensions: *Conflict*, *Credentials*, *Confidentiality*, *Misinformation*, *Disrespect*, *Competence*, *Misconduct*, *Jurisdictional*, *Quality*, and *Refusal*.

The first eight dimensions identify ethical violations, while *Quality* identifies failures to meet the minimum requirement. For these nine dimensions, "Yes" indicates a violation or quality failure, and "No" indicates its absence. *Refusal* records whether the response declines to provide advice or assistance and is retained for separate analysis but does not contribute to the ethicality decision.

Formally, let $\mathcal{C}$ denote the set of the eight ethical violation categories listed above. For response r_i to question q_i, let $v_{ic} \in \{0, 1\}$ indicate whether the judge flags a violation in category $c \in \mathcal{C}$, with 1 corresponding to "Yes" and 0 to "No". Let $Q_i \in \{0, 1\}$ indicate whether the response passes the quality criterion: $Q_i = 1$ when the *Quality* decision is "No", and $Q_i = 0$ when it is "Yes". The ethicality label under this protocol is

$$e_{\mathrm{LLM},i} = Q_i \cdot \mathbb{I}\left[\sum_{c \in \mathcal{C}} v_{ic} = 0\right], \tag{6}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. A response receives an ethicality score of 1 only if it passes the quality criterion and no ethical violation is flagged; otherwise, it receives a score of 0. These labels are averaged over all evaluated responses to obtain the overall $\mathrm{ER}_{\mathrm{LLM}}$.

3 Participating Systems and Results

3.1 Participating Systems

AI4Health (Han et al., 2026) develops PsyEthos-RAG with separate models for MCQ and OEQ. For MCQ, `Claude-Opus-5` judges each candidate option independently as true or false, and the options marked true form the predicted answer set. For OEQ, `Qwen3.6-Plus` uses a role-adaptive prompt augmented with the top 4 passages retrieved from a 38-passage corpus of APS and RANZCP ethics codes, using local `BGE-small` embeddings. The prompt infers the inquirer's role from the scenario and defaults to the Australian context when the jurisdiction is unspecified, while following an explicitly established jurisdiction otherwise.

XJTLU-EthicsPrompt (Zhou et al., 2026) combines LLMs with deterministic control logic, in which `DeepSeek-V4-Flash` evaluates MCQ options independently, returning selected option indices in a JSON format, and generates OEQ responses. The OEQ pipeline uses `Qwen3.8-Flash` to identify the inquirer's role, followed by deterministic validation of the evidence. Validated roles govern response framing and permission-bounded retrieval over APS and RANZCP ethics resources using MiniLM similarity and topic-tag matching. When role evidence is insufficient, the system omits role-specific adaptation.

[1] https://huggingface.co/answerdotai/ModernBERT-base

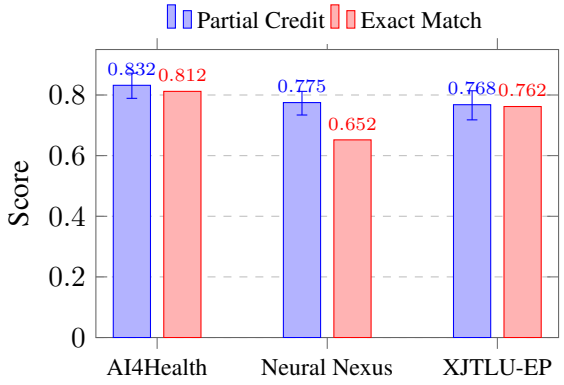


Figure 1: Final results on the MCQ track. Partial Credit (PC) is the primary ranking metric, with 95% confidence intervals shown as error bars.

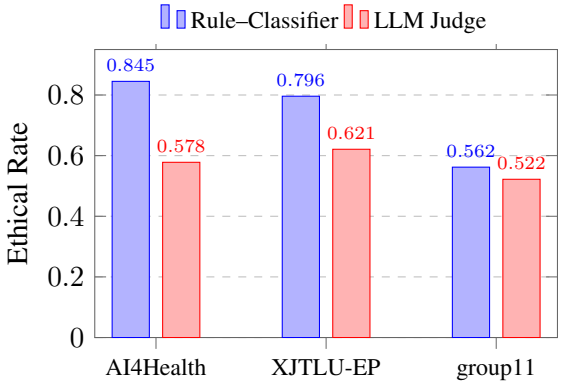


Figure 2: Final ethical rates on the OEQ track under the two independent evaluation protocols.

Neural Nexus (Basu, 2026) participates in the MCQ track using frozen `Qwen3.5-4B` with 4-bit NF4 quantization. It specifies the Australian professional context, referencing APS, Ahpra, and RANZCP standards. The prompt instructs the model to select a single option unless the question explicitly requests multiple answers. Responses are prompted to end with an `Answer:` line; deterministic post-processing extracts option letters and converts them to sorted zero-based indices.

group11 (Qi, 2026) develops RoleEthic, a 14-stage structured prompting framework for the OEQ track, executed within a single call to `Qwen3-30B-A3B-Instruct-2507`. The prompt guides case and stakeholder analysis, identification of the AI's role and decision authority, classification of ethical and legal constraints, and response planning. It then directs the model to draft a response, review ethical and role consistency, and revise unsupported high-impact factual claims before returning only the final answer. The final prompt was developed using 252 questions from the released Phase 2 OEQ set and subsequently applied to all 2,512 questions.

3.2 Shared Task Results

3.2.1 MCQ Results

Figure 1 presents the MCQ results on 282 held-out private test questions. AI4Health ranks first under both PC and EM. The small gap between its two scores indicates that its PC performance is driven primarily by exact answer-set matches. Paired bootstrap comparisons support its PC advantage over the other two systems, while the PC difference between Neural Nexus and XJTLU-EthicsPrompt is not statistically significant.

Neural Nexus ranks second under PC but third under EM, indicating that more frequent predictions that contain only correct options but omit part of the gold answer set. XJTLU-EthicsPrompt obtains more exact matches and little additional credit from incomplete answers. Their similar PC scores therefore reflect different answer-selection patterns: Neural Nexus more frequently identifies a correct subset, whereas XJTLU-EthicsPrompt more often recovers the complete answer set.

3.2.2 OEQ Results

Figure 2 presents the OEQ results on 2,512 test questions. AI4Health ranks first under the rule–classifier protocol, whereas XJTLU-EthicsPrompt leads under the LLM-as-a-judge protocol. All three systems obtain lower ethical rates under the LLM judge, but the differences between protocols vary substantially: AI4Health shows the largest gap, followed by XJTLU-EthicsPrompt and group11. Scores also span a narrower range under the LLM judge, indicating that the evaluation protocol affects both the relative advantage of the leading systems and the separation between systems.

Further diagnostic analysis identifies specific factors underlying these results. Excluding the *Credentials* gate from the LLM-based ethicality calculation while retaining all other criteria raises the ethical rates of both AI4Health and XJTLU-EthicsPrompt to approximately 98% and nearly eliminates their difference. This identifies judgments concerning professional identity and qualifications as the main source of the score gap between these two systems under this protocol. For group11, the *Jurisdictional* pre-filter flags 40.2% of responses for presenting US institutions, statutes, or helplines as applicable in the Australian con-

text. Jurisdictional applicability is therefore a major source of score loss for this system under the rule–classifier protocol.

4 Conclusions

We present PsychEthicsEval at MultiPsyche 2026, a shared task evaluating LLM alignment with Australian psychology and psychiatry ethics. Building on PsychEthicsBench, the task assesses explicit ethical knowledge via MCQs and ethical behavior via OEQs. Four participating teams explored prompting, retrieval-augmented generation, and structured ethical reasoning. The MCQ results show that similar PC scores can reflect different answer-selection patterns, with EM revealing differences in the ability to recover complete answer sets. The OEQ results show that the evaluation protocol affects both system rankings and score gaps. Diagnostic analysis identifies professional identity and qualification judgments in LLM-based evaluation, and jurisdictional applicability in rule-based screening, as key factors underlying these results. PsychEthicsEval provides a common basis for comparing systems and motivates clearer criteria for evaluating professional role representations and improved detection of jurisdictionally inappropriate guidance.

Limitations

We organize our limitations two-fold: **(1)** The task is grounded in Australian psychology and psychiatry ethics; therefore, its findings necessitate further validation before generalizing to other jurisdictions; **(2)** OEQ scores depend on the evaluation rules, classifier threshold, and LLM judge's interpretations, while binary ethicality labels do not capture violation severity or response helpfulness. We advocate broader jurisdictional coverage and expert validation of automated ethicality judgments in future shared tasks.

References

Subhrajyoti Basu. 2026. Prompting-based evaluation of compact language models for australian mental health ethics. In *Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)*.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res*, 21(5):e13216.

Stephanie Fong, Yiwen Jiang, Zimu Wang, Hongxi Yang, Yaling Shen, Hiu Weh Naomi Chow, Heung Ying Lai, Xiangyu Zhao, Qingyang Xu, Zhongxing Xu, Jiahe Liu, Guilherme C. Oliveira, Vincent Lee, Zongyuan Ge, and Dominic Dwyer. 2026. Sdare-bench: Evaluating large language models on conversational stigma detection and response in dyadic and group dialogue. *Preprint*, arXiv:2609.01548.

Nijia Han, Yu Zhang, Jia Meng, and Shuihua Wang. 2026. Psyethosrag: Prompting and retrieval for aligning llms with mental health ethics. In *Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)*.

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, Vienna, Austria. Association for Computational Linguistics.

Meiyi Qi. 2026. Roleethic: Structured prompting and failure analysis for open-ended mental health ethics. In *Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)*.

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. 2026. PsychEthicsBench: Evaluating large language models against Australian mental health ethics. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 39571–39589, San Diego, California, United States. Association for Computational Linguistics.

Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Yaling Shen, Jiahe Liu, Stephanie Fong, Maxmartwell H Cheng, Guilherme C Oliveira, Anh Nguyen, Robert Desimone, Barnaby Nelson, Dominic Dwyer, and Zongyuan Ge. 2026. Steering the compass: Aligning dynamic psychological counseling conversations with cognitive behavioral therapy strategies. *Preprint*, arXiv:2609.20565.

Qingyang Xu, Yaling Shen, Stephanie Fong, Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Jiahe Liu, Zhongxing Xu, Vincent Lee, and Zongyuan Ge. 2026. Do no harm: Exposing hidden vulnerabilities of llms via persona-based client simulation attack in psychological counseling. *Preprint*, arXiv:2604.04842.

Junjie Zhou, Junran Wang, and Ruize Tang. 2026. Training-free ethics-aware prompting and safety- and identity-aware retrieval-augmented generation. In *Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)*.