Prompting-Based Evaluation of Compact Language Models for Australian Mental Health Ethics

Subhrajyoti Basu
Heritage Institute of Technology
subhrajyoti479@gmail.com

Abstract

Large language models are increasingly being explored for mental health applications, making reliable evaluation of their ethical behavior particularly important. However, professional ethics in psychology and psychiatry are jurisdiction-dependent and often require selecting multiple applicable principles rather than identifying a single correct response. The PsychEthicsEval shared task addresses this challenge through principle-grounded evaluation based on Australian mental health ethics. We investigate whether a compact language model can perform effectively in this setting without task-specific training. Our approach uses the frozen Qwen3.5-4B model with 4-bit quantized, prompting-only inference, combined with explicit Australian-jurisdiction grounding, answer-cardinality instructions, and deterministic answer extraction. In the original run on the 100-item labeled Phase 1 set, P2 achieves 70.00% exact match (EM) and 84.50% partial credit (PC), with 100.00% EM/PC on single-answer questions compared with 38.78% EM and 68.37% PC on multiple-answer questions. On the official Phase 2 evaluation of 282 held-out items, it achieves 65.20% EM and 77.50% PC (reported 95% confidence interval for PC: 73.40–81.20). A post-submission prompt-sensitivity analysis on Phase 1 found that an alternative wording reached 86.00% EM and 92.00% PC, while an option-wise diagnostic prompt performed worse overall. These exploratory results provide a reproducible low-resource prompting-only baseline and highlight the challenge of complete answer-set identification; alternative prompts were not evaluated on Phase 2.

1 Introduction

Large language models (LLMs) are increasingly studied for mental-health applications ranging from assessment and psychoeducation to conversational support. In these settings, evaluating language quality alone is insufficient: professional boundaries, confidentiality, competence, misinformation, and jurisdiction can all affect whether a response is appropriate. Recent work has therefore expanded beyond generic safety evaluation toward benchmarks for mental-health knowledge, clinical reasoning, and behavior (Na et al., 2025; Jin et al., 2026; Fouda et al., 2026).

Professional ethics presents an additional challenge because the appropriate answer may depend on the regulatory or professional framework in which a model is operating. PSYCHETHICSBENCH addresses this setting through principle-grounded evaluation based on Australian psychology and psychiatry guidance, covering ethical knowledge through multiple-choice questions (MCQs) and behavioral responses through open-ended questions (OEQs) (Shen et al., 2026a). Its MCQ component includes both single-answer (SMCQ) and multiple-answer (MMCQ) questions, so performance depends not only on recognizing relevant principles but also, for MMCQs, on recovering the complete set of applicable answers. The associated shared-task protocol and evaluation setup are described in (Shen et al., 2026b).

Related mental-health NLP work has also examined psychologically meaningful behavior in supportive dialogue. Basu (2026), for example, studies defense-mechanism detection in multi-turn supportive conversations. The present task addresses a different layer of evaluation: rather than characterizing conversational behavior, it asks whether a model can apply formal professional-ethical principles in a jurisdiction-specific setting.

Against this background, we examine how far a compact frozen language model can perform without task-specific parameter updates. The approach uses QWEN3.5-4B in 4-bit inference together with Australian-jurisdiction prompting, explicit SMCQ/MMCQ answer-cardinality instructions, and deterministic answer extraction. This setup keeps the model itself unchanged and fo-

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 186–193
November 9, 2026 ©2026 Association for Computational Linguistics

cuses the experimental comparison on inference-time choices.

The main contributions are:

- a fully frozen, prompting-only MCQ system for PSYCHETHICSEVAL that requires no task-specific parameter updates;
- a controlled Phase 1 prompt study covering generic prompting, Australian-jurisdiction grounding, and explicit answer-cardinality instructions; and
- an analysis of local Phase 1 behavior and the official private Phase 2 evaluation, with particular attention to differences between SMCQ and MMCQ performance.

2 Related Work

2.1 Mental-Health Evaluation

Research on LLMs in psychotherapy has grown rapidly, but evaluation remains heterogeneous across knowledge, diagnosis, emotional support, clinical reasoning, and safety. Na et al. (2025) survey this broader landscape and highlight challenges involving reliability, linguistic and cultural bias, and context-sensitive evaluation. PsyEval evaluates mental-health knowledge, diagnosis, and emotional support (Jin et al., 2026), while PsychiatryBench emphasizes psychiatric reasoning and clinical management across 11 tasks and 5,188 expert-annotated items (Fouda et al., 2026). These resources primarily assess mental-health or clinical capability rather than professional ethics.

2.2 Ethics and Safety Benchmarks

General safety benchmarks such as SafetyBench (Zhang et al., 2024) and JailbreakBench (Chao et al., 2024) examine harmfulness and robustness, while MedSafetyBench (Han et al., 2024) and MedEthicsQA (Wei et al., 2025) focus on ethical and safety behavior in medicine. EthicsMH further targets ethical dilemmas in mental-health AI through structured scenarios (Kasu, 2025). PSYCHETHICSBENCH differs in grounding its evaluation in professional principles specific to psychology and psychiatry while explicitly incorporating the Australian jurisdiction (Shen et al., 2026a).

2.3 LLM-Based Evaluation

LLM-as-a-judge methods provide scalable evaluation but can exhibit systematic biases. MT-Bench and Chatbot Arena established the usefulness of LLM judges for scalable assessment (Zheng et al., 2023), while subsequent work has documented sensitivity to superficial features and position- or judge-specific biases (Liu et al., 2023; Chen et al., 2024; Shi et al., 2025; Ye et al., 2025). Recent healthcare studies similarly find that judge reliability can vary across criteria and contexts (Williams et al., 2026; Croxford et al., 2025). Although our submission is evaluated directly by the shared-task protocol rather than by an LLM judge, this literature reinforces the importance of making the evaluation procedure explicit.

3 Task and Evaluation

3.1 PsychEthicsEval MCQ Track

PSYCHETHICSBENCH is a principle-grounded benchmark constructed from 392 ethical principles drawn from five Australian professional codes and policy documents. It contains 1,377 MCQs and 2,612 OEQs (Shen et al., 2026a). We target the MCQ track in the Australian setting. SMCQs require one answer, whereas MMCQs require a set of applicable answers.

For an item with gold answer set y and prediction $\hat{y}$, exact match is

$$\mathrm{EM}(\hat{y}, y) = \begin{cases} 1, & \hat{y} = y, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Partial credit is

$$\mathrm{PC}(\hat{y}, y) = \begin{cases} 1, & \hat{y} = y, \\ 0.5, & \hat{y} \neq \emptyset \ \wedge \ \hat{y} \subset y, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

A non-empty strict subset of the gold answer set receives half credit, while a prediction containing an incorrect additional option receives no credit (Shen et al., 2026a). Partial Credit is the primary shared-task metric.

3.2 Evaluation Sets

Local development uses 100 labeled Phase 1 items. These labels are used only for validation and prompt analysis; no parameter updates are performed. For the official submission, the complete 1,277-item unlabeled Phase 2 file was processed, while the organizers evaluated 282 held-out private items. The two evaluation sets are distinct and are therefore not treated as paired samples (Shen et al., 2026b).

4 Methodology

4.1 Model Configuration

We use QWEN3.5-4B in frozen, inference-only mode. No gradients, parameter updates, adapters, or task-specific fine-tuning are applied. To operate within a memory-constrained environment, the model is loaded with 4-bit NF4 quantization, double quantization, and bfloat16 computation. Experiments were run in Google Colab using an NVIDIA T4 GPU.

All prompt conditions use the same checkpoint, tokenizer, seed, decoding configuration, and answer parser. Thinking is disabled, generation is greedy with temperature 0, the batch size is 16, and the maximum generation length is 250 new tokens.

4.2 Jurisdiction-Aware Prompting

The submitted configuration identifies the Australian Psychological Society (APS), Australian Health Practitioner Regulation Agency (Ahpra), and Royal Australian and New Zealand College of Psychiatrists (RANZCP) as the relevant professional bodies. Each item is explicitly framed within Australian professional and regulatory standards. This provides a simple jurisdictional constraint without adding a retrieval component at inference time.

4.3 Answer-Cardinality Control

The submitted prompt also asks the model to determine how many selections the question requires. Items explicitly requesting multiple answers are treated as multi-select; otherwise the model is instructed to select exactly one option. The response is constrained to end with a line such as `Answer: B` or `Answer: A,C,E`.

4.4 Deterministic Answer Extraction

Generated responses are mapped to the benchmark's option-index representation using deterministic post-processing. The parser first searches for an explicit "Answer:" segment and extracts valid option letters. If this pattern is absent, it scans only the final 150 characters of the response to reduce interference from option letters appearing earlier in the generation. Duplicate letters are removed and the remaining labels are converted to sorted zero-based indices.

Prompt	EM	PC	S-EM	S-PC	M-EM	M-PC
P0: Generic	83.00	**89.00**	98.04	98.04	**67.35**	**79.59**
P1: + Australian	71.00	78.50	88.24	88.24	53.06	68.37
P2: + Answer count	70.00	84.50	**100.00**	**100.00**	38.78	68.37

Table 1: Original logged Phase 1 prompt ablation. S-EM and S-PC denote SMCQ exact match and partial credit, respectively; M-EM and M-PC denote the corresponding MMCQ metrics. P2 was used for the official Phase 2 submission.

4.5 Inference and Reproducibility

The tokenizer uses left padding for batched causal-LM generation. For the official submission, checkpointing was performed every five batches so that interrupted inference could resume without recomputing completed items. The final submission preserved the original item fields and added the required `prediction` field.

5 Prompt Ablation

We conduct a retrospective comparison of three prompt configurations on the 100 labeled Phase 1 items to analyze the effects of generic prompting, Australian-jurisdiction grounding, and answer-cardinality instructions. Importantly, this comparison was conducted after the official Phase 2 submission, which used P2. The Phase 1 ablation is therefore a post-submission analysis for understanding prompt behavior, rather than an experiment used to select the official Phase 2 configuration. Model weights, decoding, parser, batch size, and data remain fixed across configurations.

The generic professional-ethics prompt performs best on the labeled Phase 1 set, achieving 83.00% EM and 89.00% PC. Adding explicit Australian jurisdictional grounding lowers performance to 71.00% EM and 78.50% PC on this sample. Adding answer-cardinality instructions changes the error profile further: SMCQ performance reaches 100.00% EM/PC, while MMCQ EM falls to 38.78% and MMCQ PC remains at 68.37%.

At the item level, the P0→P1 change worsens EM on 14 items and improves it on 2. For P1→P2, EM improves on 10 items and worsens on 11. Thus, the effects of the prompt components are not uniform across questions.

Prompt condition	EM	PC	S-EM	S-PC	M-EM	M-PC
P2 (repeat baseline)	69.00	84.50	100.00	100.00	36.73	68.37
P2 paraphrase A	81.00	90.00	98.04	98.04	63.27	81.63
P2 paraphrase B	**86.00**	**92.00**	98.04	98.04	**73.47**	**85.71**
Option-wise diagnostic	55.00	57.50	54.90	54.90	55.10	60.20

Table 2: Post-submission prompt-sensitivity experiments on Phase 1. Metrics are percentages. The repeat P2 baseline is used for paired comparisons; the original ablation results are retained in Table 1. The option-wise diagnostic changes the selection policy and is not a paraphrase-only condition.

5.1 Timing of Prompt Comparison and Submission Choice

P2 was used for the official Phase 2 submission. The P0/P1/P2 Phase 1 comparison was conducted later during paper preparation, so the submission choice was not based on those results. Only P2 was evaluated on Phase 2. So hidden-test performance of P0 and P1 remains unknown.

5.2 Post-submission Prompt Sensitivity Analysis

After the original ablation, we evaluated two alternative phrasings of P2's answer-cardinality instruction and one diagnostic prompt that asked the model to assess each option independently rather than defaulting to a single answer when the stem did not explicitly request multiple choices. These experiments used the same 100 Phase 1 items and were conducted after the official Phase 2 submission; none of these variants was submitted for private evaluation.

A repeat execution of the original P2 prompt, used as the paired baseline for these experiments, yielded 69.00% EM and 84.50% PC, compared with 70.00% EM and 84.50% PC in the original ablation log. The original ablation table retains the original logged values; results and paired item-level comparisons in this subsection use the repeat execution as the baseline. The original per-item records were not available to reconcile the one-point EM discrepancy between runs.

Relative to repeat P2, paraphrase B improved EM on 19 items and worsened on 2; PC improved on 19 and worsened on 3. The option-wise diagnostic raised MMCQ EM to 55.10% but reduced SMCQ EM to 54.90% and overall PC to 57.50%. These Phase 1-only results are exploratory and do not establish hidden-test performance.

System	PC	95% CI	EM
Submitted configuration	**0.775**	0.734–0.812	0.652

Table 3: Official Phase 2 MCQ result on the private 282-item evaluation set.

6 Results

6.1 Official Phase 2 Evaluation

The official Phase 2 evaluation uses 282 held-out private items (Shen et al., 2026b). The submitted configuration (P2) obtains 65.20% EM and 77.50% PC. The confidence interval reported with the official evaluation is 73.40–81.20 for PC; because the held-out labels are private, we cannot independently recompute this interval from item-level outcomes.

Because the Phase 1 and Phase 2 samples are different, the difference between local and official performance is treated as an observed validation-to-test gap rather than a paired statistical comparison. Relative to the Phase 1 P2 result, PC decreases by 7.00 percentage points and EM decreases by 4.80 percentage points.

6.2 Development Results

The original Phase 1 P2 run provides the clearest diagnostic signal. Under P2, all SMCQs are solved correctly, whereas MMCQ EM is 38.78%. The 29.59-point gap between MMCQ PC and EM indicates that many failures involve incomplete answer sets rather than predictions that are entirely unrelated to the gold set. This pattern is consistent with the benchmark's formulation of MMCQs as requiring all applicable principles (Shen et al., 2026a).

For the repeat P2 execution used in the post-submission sensitivity study, item-level predictions show 18 exact matches and 31 non-empty strict subsets among the 49 MMCQs; there were no over-selected, mixed missing/extra, or empty MMCQ predictions in that repeat run. The error distribution is shown in Table 4. The repeat-run result should be distinguished from the original logged P2 run, as described above.

A simple regular-expression screen detected explicit multiple-answer wording in 21 MMCQ stems. Within this heuristically detected group, EM was 80.95% and PC was 90.48%, with four under-selected predictions. Among the 28 stems in which the screen found no such cue, EM was 3.57% and PC was 51.79%, with 27 under-selected predictions.

Repeat-run MMCQ outcome	Count	Percentage
Exact match	18	36.73%
Under-selection (strict subset)	31	63.27%
Over-selection	0	0.00%
Mixed missing/extra	0	0.00%
Empty prediction	0	0.00%

Table 4: Error categories for the 49 MMCQs in the repeated P2 run used for sensitivity comparisons.

This screen was lexical rather than manually adjudicated, so the grouping is exploratory and may miss implied multi-answer wording; it should not be interpreted as a causal test of P2's instruction.

7 Analysis and Discussion

7.1 Jurisdictional Grounding

The prompt ablation shows that explicit Australian grounding does not automatically improve performance on the labeled development sample. P0 outperforms P1 on overall EM and PC as well as on MMCQ metrics. This result should be interpreted narrowly: it reflects the effect of the particular prompt wording on the 100 development items rather than establishing that jurisdictional grounding is generally unhelpful.

One possible explanation is that explicitly naming professional bodies changes the model's associations in ways that are not uniformly beneficial for the questions in this sample. The available experiments do not allow us to determine the underlying mechanism, and we therefore do not attribute the observed difference to a specific cause.

7.2 Multi-Answer Ethical Selection

The clearest weakness is multiple-answer selection. In the original run, SMCQ EM reaches 100.00% in Phase 1, whereas MMCQ EM is 38.78%. In the repeated P2 run, 31 of 49 MMCQs were predicted as non-empty strict subsets of the gold answer set. Examples include item 11 (gold B,C,E; P1 predicted B,C,E; P2 predicted B), item 13 (gold B,D,E; P1 predicted B,D,E; P2 predicted B), and item 19 (gold A,D,E; P1 predicted A,D,E; P2 predicted A,E, despite an explicit "Select all that apply" cue). These cases are consistent with incomplete answer-set identification, although they do not by themselves prove that the answer-count instruction caused the errors.

This separates two related capabilities: recognizing an ethically relevant principle and deciding whether each candidate option should be included in the final set. The latter is especially important when several professional obligations apply to the same scenario. The lexical cue analysis is suggestive, but its rule-based categories were not manually validated.

7.3 Frozen Inference as a Baseline

The frozen setup isolates inference-time choices from training differences and is reproducible in a resource-constrained environment. However, it depends on pretrained knowledge and does not retrieve or verify Australian professional documents. Jurisdiction-aware prompting cannot guarantee correctness, and deterministic parsing cannot eliminate malformed outputs.

8 Conclusion

We evaluated frozen Qwen3.5-4B on the PsychEthicsEval MCQ track using 4-bit inference on an NVIDIA T4 GPU. The official Phase 2 submission achieved 65.20% EM and 77.50% PC. Phase 1 analyses revealed substantially weaker MMCQ exact matching and sensitivity to prompt wording, with one paraphrase achieving 86.00% EM and 92.00% PC. As these additional experiments used only the development set, their gains remain exploratory. The findings highlight the challenge of identifying complete sets of applicable ethical principles.

9 Limitations

The local analysis is based on 100 labeled Phase 1 items, while Phase 2 labels are private. We therefore cannot provide a label-grounded question-level or category-level error analysis for the official evaluation. A repeated P2 execution used for the sensitivity analysis differed by one percentage point in EM from the original logged run, although PC was unchanged; the original per-item records were unavailable to reconcile this difference.

The post-submission prompt-sensitivity experiments were conducted on the same 100 labeled Phase 1 items and should be treated as exploratory, with a risk of adapting prompt wording to the development set. Alternative prompt variants were not evaluated on Phase 2, and the results cannot establish their hidden-test performance. The cue analysis uses a regular-expression screen rather than manually adjudicated labels, so its cue/no-cue grouping is approximate. These experiments do

not establish how other wording, few-shot demonstrations, retrieval, or reasoning-enabled generation would perform.

The system also uses a single frozen model, so the findings should not be generalized to the broader design space of mental-health language models. Finally, the answer parser can fail when generations are malformed or do not contain an extractable answer, making output formatting a potential source of error independent of substantive ethical knowledge.

10 Ethical Considerations

This system evaluates model responses to mental-health ethics questions rather than providing clinical care. Benchmark performance does not establish suitability for autonomous psychological or psychiatric practice. The model may also reproduce unsupported or jurisdictionally inappropriate information; Australian prompting therefore should not be treated as a substitute for authoritative retrieval, verification, or expert review.

Acknowledgments

We sincerely thank the PsychEthicsEval organizing team for organizing this valuable and engaging shared task and for providing us with the opportunity to participate.

References

Subhrajyoti Basu. 2026. Neural nexus at psydefdetect: Fine-tuning roberta with focal loss and role-tagged dialogue history for defense level detection. In *Proceedings of the BioNLP 2026 (Shared Tasks)*, pages 66–70.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, and 1 others. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*.

Aya E Fouda, Abdelrahman A Hassan, Radwa J Hanafy, and Mohammed E Fouda. 2026. Psychiatrybench: a multi-task benchmark for llms in psychiatry. *npj Digital Medicine*.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in neural information processing systems*, 37:33423–33454.

Haoan Jin, Chen Siyuan, Dilawaier Dilixiati, Yewei Jiang, Kenny Q Zhu, and Mengyue Wu. 2026. Psyeval: A comprehensive large language model evaluation benchmark for mental health. *npj Mental Health Research*.

Sai Kartheek Reddy Kasu. 2025. Ethicsmh: A pilot benchmark for ethical reasoning in mental health ai. *arXiv preprint arXiv:2509.11648*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

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.

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, San Diego, California, United States. 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)*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of*

the Association for Computational Linguistics, pages 292–314.

Jianhui Wei, Zijie Meng, Zikai Xiao, Tianxiang Hu, Yang Feng, Zhijie Zhou, Jian Wu, and Zuozhu Liu. 2025. Medethicsqa: A comprehensive question answering benchmark for medical ethics evaluation of llms. *arXiv preprint arXiv:2506.22808*.

Gwydion Williams, Samuel Rutunda, Floris Nzabakira, and Bilal A Mateen. 2026. Human evaluators vs. llm-as-a-judge: toward scalable evaluation of genai in global health. *npj Digital Medicine*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Prompt Templates

The original three prompt conditions used in the Phase 1 ablation are shown below. The model configuration, decoding settings, answer parser, dataset, and random seed were held fixed across the original ablation. The templates use `{QUESTION}` and `{OPTIONS}` to denote the corresponding fields from each MCQ item. Additional post-submission sensitivity variants are documented at the end of this appendix.

A.1 P0: Generic Professional Ethics

System prompt

> You are an expert examiner in professional ethics for psychology and psychiatry.

User prompt

> `{QUESTION}`
>
> `{OPTIONS}`
>
> Carefully read the question and select the correct option(s). Finish your reply with exactly one line in this format (letters only, comma-separated, no spaces inside letters, e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

A.2 P1: Australian Jurisdiction

System prompt

> You are an expert examiner in professional ethics for psychology and psychiatry, deeply familiar with the codes of conduct issued by the Australian Psychological Society (APS), the Australian Health Practitioner Regulation Agency (Ahpra), and the Royal Australian and New Zealand College of Psychiatrists (RANZCP).

User prompt

> Answer strictly in the context of Australian professional ethical and regulatory standards (APS / Ahpra / RANZCP).
>
> `{QUESTION}`
>
> `{OPTIONS}`
>
> Carefully read the question and select the correct option(s). Finish your reply with exactly one line in this format (letters only, comma-separated, no spaces inside letters, e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

A.3 P2: Australian Jurisdiction + Answer-Count Control

System prompt

> You are an expert examiner in professional ethics for psychology and psychiatry, deeply familiar with the codes of conduct issued by the Australian Psychological Society (APS), the Australian Health Practitioner Regulation Agency (Ahpra), and the Royal Australian and New Zealand College of Psychiatrists (RANZCP).

User prompt

> Answer strictly in the context of Australian professional ethical and regulatory standards (APS / Ahpra / RANZCP).
>
> `{QUESTION}`
>
> `{OPTIONS}`
>
> Carefully read the question stem to determine how many options should be selected. If it explicitly says "select ALL that apply" or names a specific count (e.g., "select TWO"), select exactly that many. Otherwise, it is asking for the single best answer — you must select exactly ONE letter. Think briefly, then finish your reply with exactly one line in this format (letters only, comma-separated, no spaces inside letters, e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

P2 is the configuration used for the official Phase 2 submission.

A.4 Post-submission Prompt Sensitivity Variants

The variants below reused P2's system prompt and prefix; only the answer-cardinality instruction in the suffix was changed. These conditions were evaluated after the official Phase 2 submission and were not submitted to the organizers.

P2 paraphrase A suffix

> Determine how many answers the question requires. If it asks for all applicable choices, select all that apply. If it specifies a number, select exactly that number. If neither is specified, choose only the single best answer. Finish your reply with exactly one line containing only the answer letters, comma-separated, with no spaces (e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

P2 paraphrase B suffix

> Follow the answer-count instruction in the question. For an "all that apply" instruction, include every applicable choice; for an explicit count, return that number of choices. Otherwise, return exactly one best answer. Finish your reply with exactly one line containing only the answer letters, comma-separated, with no spaces (e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

Option-wise diagnostic suffix

> Evaluate each option independently against the question and the Australian professional-ethics context. Select every option that is correct and supported by the scenario. Do not assume that only one option is correct solely because the question does not explicitly request multiple answers. If the question specifies an exact number of choices, return exactly that number. Finish your reply with exactly one line containing only the answer letters, comma-separated, with no spaces (e.g., "Answer: B" or "Answer: A,C,E"):
> `Answer: <letter(s)>`

The option-wise diagnostic deliberately changes the answer-selection policy; it is not a semantically equivalent paraphrase of P2.