

VAR-Safe: Safety-Gated Variational Alignment for Chinese digital psychological counseling

Xinyu Song^{1}, Zhengjie Gao¹*

¹School of Electronic Information Engineering, Geely University of China, Chengdu, China

*Corresponding Author. Email: songxinyu@guc.edu.cn

Abstract. Large Language Models (LLMs) show immense potential in Chinese digital psychological counseling services. However, their training alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF), face challenges including implementation complexity, high computational cost, and training instability. These issues are particularly critical in the high-safety-requirement context of psychological counseling, where model Hallucination and ethical risks urgently need to be addressed. Guided by the safety-first principle, this paper proposes the Safety-Gated Variational Alignment (VAR-Safe) method, built upon the foundation of the Variational Alignment (VAR) technique. VAR-Safe introduces a safety-gated reward transformation mechanism that converts the professional ethics and harmlessness constraints encoded in the reward model into hard penalty terms, thereby more effectively suppressing harmful or unprofessional hallucinated responses. From the perspective of variational inference, VAR-Safe transforms the complex objective of RLHF into an offline, safety-driven, re-weighted Supervised Fine-Tuning (SFT) format. This ensures that all weights during the optimization process remain positive, fundamentally enhancing the robustness and convergence stability of the alignment training. We trained a Chinese digital psychological counselor based on the Chinese SoulChat corpus. Experimental results show that while significantly improving the model's empathy and professionalism, VAR-Safe reduces the critical safety metric—the rate of professional knowledge hallucination—to a level much lower than that of the baseline models, demonstrating its superior applicability in high-safety applications.

Keywords: LLMs, psychological counseling, reinforcement learning

1. Introduction

The capabilities of Large Language Models (LLMs) make them an ideal foundation for constructing Chinese digital psychological counselors. This service not only offers timely psychological support but also aligns with the current societal demand for convenient and accessible mental health services. However, psychological counseling is a high-stakes, high-ethical-requirement domain. The content generated by the model must be accurate and professional; any false (i.e., hallucinated) or unprofessional advice could potentially harm the user. According to the New Generation Artificial Intelligence Ethics Norms in China, AI applications must adhere to fundamental ethical principles such as safety and transparency, and enhanced risk prevention. Therefore, the core technical challenge lies in how to efficiently and stably integrate these stringent ethical and safety standards into the LLM's generation strategy.

The classic Reinforcement Learning from Human Feedback (RLHF) [1] typically employs algorithms such as Proximal Policy Optimization (PPO) [2]. Although effective, PPO is complex to implement, computationally intensive, and its policy clipping mechanism may introduce gradient bias. To simplify this process, the research community has proposed offline alignment methods like Direct Preference Optimization (DPO) [3]. However, these simplified methods, when dealing with extremely high or low reward samples, can still suffer from unstable gradients or experience "synchronous likelihood collapse" due to negative log ratios, thereby compromising model performance. Such training instability is unacceptable in the high-safety-requirement context of counseling [4].

Variational Alignment (VAR) was introduced to address these issues. Theoretically, it transforms the RLHF optimization problem into minimizing the distribution gap between the learned policy and the optimal policy, which is ultimately formalized as a reward-driven re-weighted Supervised Fine-Tuning (SFT) objective. The key advantage of VAR is that its constructed weights are guaranteed to be positive, thereby ensuring the robustness of the optimization process.

While VAR provides superior stability and efficiency, in the unique, high-risk scenario of psychological counseling, the model must possess an even greater sensitivity to professional ethical violations and factual hallucinations. A simple reward score may be insufficient to impose adequate penalty on these fatal errors, especially in the early stages of training [5].

This paper proposes the Safety-Gated Variational Alignment (VAR-Safe) method. Our main contributions are as follows:

1. **Introducing the VAR-Safe Framework:** Building upon the core mechanism of VAR, we introduce a safety-gated reward transformation mechanism specifically tailored for psychological counseling. By identifying and penalizing responses that violate the Chinese Psychological Society Code of Ethics for Clinical and Counseling Psychology Professionals or constitute professional knowledge hallucination, this mechanism elevates compliance with safety ethics from a soft constraint to a hard penalty, thus significantly enhancing hallucination suppression capabilities.

2. **Reinforcing Ethical Safety Alignment:** VAR-Safe ensures, via the safety gate, that responses with low safety scores are assigned extremely low weights during the re-weighted SFT process. This allows the policy to converge faster to a high-safety, low-hallucination state that conforms to professional ethical norms.

3. **Efficient Implementation and Validation:** We leverage the efficient In-Batch Normalization technique from VAR for scalable training of VAR-Safe on the Chinese SoulChat corpus. Experiments confirm that VAR-Safe surpasses baseline methods like PPO and DPO in terms of hallucination rate, harmlessness, and professionalism when aligning Chinese psychological counseling LLMs.

2. Related work

2.1. RLHF and simplified alignment based on variational inference

Traditional Reinforcement Learning from Human Feedback (RLHF) relies on policy gradients and online sampling, the high cost of which has prompted researchers to explore offline, off-policy simplification methods. Direct Preference Optimization (DPO) successfully transforms the RLHF objective into a simple cross-entropy loss by establishing a relationship between the policy ratio and the reward function. However, DPO is prone to gradient explosion or convergence difficulties when dealing with high-variance data.

The Variational Alignment (VAR) method differs fundamentally from DPO. It is based on the Maximum Entropy RL framework and aims to directly approximate the closed-form solution of the optimal policy, which resembles a policy distribution with an exponential reward term [6]. Through variational inference, minimizing the gap between the learned policy and this optimal solution ultimately derives a reward-weighted negative log-likelihood loss. The elegance of this framework lies in the fact that its weighting term w is always constructed based on a positive measure, thereby avoiding the instability issues caused by negative log ratios in DPO. The theoretical foundation of VAR is highly consistent with the trend in offline reinforcement learning of transforming RL into Reweighted Supervised Learning (RWSL).

2.2. LLM application in psychological counseling

In recent years, Large Language Models (LLMs), with their robust natural language understanding and generation capabilities, have demonstrated immense potential in the field of mental health support. Researchers are exploring the use of LLMs to build digital companionship and preliminary consultation systems aimed at providing instant, convenient, and accessible emotional support and psychological guidance [7, 8]. These systems have achieved positive progress in handling general psychological distress and offering foundational exercises for Cognitive Behavioral Therapy (CBT), while also showing potential in enhancing user empathy [9].

However, as a highly sensitive and high-risk professional domain, the application of LLMs in psychological counseling faces severe ethical and safety challenges [10]. The core issue lies in the models' propensity for Hallucination—generating false, unprofessional, or potentially harmful medical or psychological advice. This hallucinatory behavior directly violates the professional psychological counseling ethical principle of "Do No Harm," potentially causing irreversible damage to users in a vulnerable state [11]. Furthermore, the intervention capabilities and ethical boundaries of LLMs in complex crisis situations, such as those involving self-harm or suicide, remain an unresolved challenge.

3. Methodology

3.1. VAR Alignment

The Variational Alignment (VAR) method aims to transform the complex RLHF alignment process into an efficient and stable offline optimization format. Based on variational inference, the core idea is to directly minimize the distributional gap between the learned policy π_θ and the optimal RLHF policy π^* . Through theoretical mathematical transformation, this objective is

equivalent to maximizing a reward-driven Weighted Supervised Fine-Tuning (Weighted SFT) Loss. In this loss formulation, model training no longer relies on complex policy gradient iterations, but is achieved by assigning a non-negative weight $w(x, y)$ to each response in the training data:

$$w(x, y) \propto \pi_{ref}(y|x) \exp(r(x, y)/\lambda) \quad (1)$$

where the weight w is exponentially and positively correlated with the reward r obtained by the response. High-reward responses are assigned greater weights, thus dominating the training process and guiding the model policy toward the optimal direction. Most importantly, this weighting term w is always non-negative, ensuring that the optimization of the loss function is a stable and unambiguous likelihood maximization process. This effectively avoids the gradient instability issues, which can be caused by negative weights, encountered in other simplified methods (such as DPO).

3.2. VAR-Safe: Safety-Gated Reward Transformation Mechanism

In the application scenario of psychological counseling, which has extremely high professional and ethical requirements, we propose a crucial original improvement to the VAR method—the Safety-Gated Reward Transformation Mechanism. While the basic reward model $r(x, y)$ can score the overall quality of a response, merely assigning a "low score" (which might still be positive) for behaviors like Factual Hallucination or violations of the Code of Ethics is insufficient to thoroughly suppress them during training. In a safety-critical context where a zero-tolerance policy is required, any unsafe or unprofessional response must be assigned an infinitely low priority.

Therefore, we introduce a safety-gated indicator function $S(x, y)$ and an extremely large, hard penalty coefficient α , defining the Safety-Enhanced Reward $r'_{safe}(x, y)$ as:

$$r'_{safe}(x, y) = r(x, y) - \alpha S(x, y) \quad (2)$$

Definition of the Safety-Gated Indicator $S(x, y)$:

$S = 1$: When response y is marked by a professional audit as containing professional knowledge hallucination (e.g., providing incorrect diagnoses or medication advice), encouraging dangerous behavior (such as self-harm, suicide, or harming others), severely violating confidentiality.

$S = 0$: Response y does not contain any of the above high-risk ethical or factual errors.

The penalty coefficient α is set as a constant much larger than the maximum value of the basic reward ($\alpha \gg \max(r)$). This ensures that once $S = 1$ (i.e., a safety violation or hallucination occurs), the enhanced reward $r'_{safe}(x, y)$ will be an extremely large negative number.

This safety-gated mechanism perfectly integrates with the exponential re-weighting property of VAR, yielding the following key safety benefits:

Exponential Hallucination Suppression: When $r'_{safe}(x, y)$ becomes an extremely large negative value due to the safety gate, the weight w will exponentially approach zero. This means samples containing professional hallucination or ethical risk lose almost all weight in the training, unable to contribute effective likelihood gradients to the policy π_θ . This achieves rapid and absolute suppression of unsafe behaviors.

Maintaining Training Stability: Although $r'_{safe}(x, y)$ is an extremely large negative number, the weight w remains non-negative through the exponential function (merely approaching zero). This maintains the inherent stability of the VAR framework, guaranteeing the robustness of the training process.

Through this method, which combines hard ethical constraints with the stability of variational alignment, VAR-Safe ensures that the trained Chinese digital psychological counselor can efficiently and stably converge to a state of high safety and low hallucination.

4. Implementation

4.1. Base model and data preparation

Base Model: The Qwen2.5-7B model was selected as the foundational model (base model) for the Chinese digital psychological counselor [12]. The Qwen2.5 model is based on the Transformer architecture, featuring 28 layers, utilizing RoPE position encoding, the SwiGLU activation function, and RMSNorm, with a total parameter count of approximately 7.61 billion (non-embedding parameters being about 6.53 billion). This model was chosen for its excellent performance on Chinese pre-training corpora and its broad compatibility with general alignment techniques.

SoulChat Corpus [13]: The data used for alignment training originated from the Chinese SoulChat corpus, which underwent strict anonymization and ethical desensitization. This corpus covers a rich array of psychological distress scenarios and professional responses, providing high-quality domain-specific data for training both the Reward Model (RM) and the Safety Gate function.

Reward Model (RM) and Safety Gate (S): The RM was trained following professional psychological counseling standards, focusing on assessing:

1. Empathy and friendliness.
2. Professionalism and harmlessness of the response.
3. Factual consistency/Hallucination risk (used to construct $S(x, y)$).

The RM provides scores for the basic reward $r(x, y)$. Concurrently, an independent professional evaluation team was responsible for conducting a safety audit of the corpus, marking all samples involving professional knowledge hallucination or ethical violations to construct the $S(x, y)$ indicator function.

4.2. Hyperparameter

Table 1 summarizes the key hyperparameters used for training VAR-Safe, including the Regularization Coefficient λ set to 1.0, the large Safety Penalty Coefficient α set to 50.0 to rigorously enforce the zero-tolerance ethical constraints, and a learning schedule defined by a Learning Rate of 10^{-6} and 3 epochs.

Table 1. Hyperparameter

Hyperparameter	Value
Regularization Coefficient λ	1.0
Safety Penalty Coefficient α	50.0
Batch Size	128
Learning Rate	10^{-6}
Epochs	3

5. Experiments and results analysis

5.1. Experimental setup and evaluation metrics

Baseline Models: To comprehensively evaluate the performance of VAR-Safe, we selected the following models as comparison baselines: the basic Supervised Fine-Tuning model (SFT Baseline), the classic Reinforcement Learning method PPO-RLHF, the popular offline alignment method Direct Preference Optimization (DPO), and the original Variational Alignment method (VAR Baseline).

Evaluation Metrics: The experiments quantified the models' alignment efficacy across multiple dimensions:

Alignment Performance: We used Human Evaluation by domain experts and advanced LLM-as-a-Judge scoring to assess Helpfulness (\uparrow) and Harmlessness (\uparrow).

Safety and Ethics: The core safety metric is Factual Consistency, quantified as the Hallucination Rate (\downarrow)—the frequency of providing incorrect professional knowledge or logical fallacies in response to professional counseling queries. We also evaluated Professional Ethical Compliance, which measures the strict adherence of responses to psychological counseling ethical norms.

Training Efficiency and Stability: We quantified the GPU-Hours required for training and the variance of the loss function or KL divergence during the training process to assess the robustness of the optimization.

5.2. VAR-Safe alignment performance and stability comparison

VAR-Safe achieves ultra-high training stability through its safety-gated and positive weighting mechanisms. The results in Table 2 validate VAR-Safe's superiority in training robustness and performance.

Table 2. Evaluation of model alignment capability and stability: comparison of VAR-Safe with mainstream baselines

Model	Helpfulness Score (\uparrow)	Harmlessness Score (\uparrow)	Training Stability (KL Variance $\times 10^{-2}$ \downarrow)	GPU Hours (\downarrow)
SFT Baseline	7.5	8.0	N/A	50
PPO-RLHF	8.5	8.2	4.5	450
DPO	8.8	8.6	3.2	150
VAR Baseline	9.0	8.9	1.5	130
VAR-Safe (Ours)	9.2	9.1	1.2	120

VAR-Safe achieved the highest scores in both Helpfulness (9.2) and Harmlessness (9.1) comprehensive evaluations. The original VAR Baseline significantly outperformed PPO and DPO in both stability and performance, reflecting the robustness derived from its non-negative weight construction. Crucially, VAR-Safe's Training Stability metric (KL divergence variance) is only 1.2×10^{-2} , significantly lower than all baselines. This confirms that VAR-Safe's safety-gated mechanism, combined with the non-negative weight principle of VAR, results in exceptionally strong intrinsic robustness. In terms of computational efficiency, VAR-Safe (120 GPU Hours) is marginally superior to the VAR Baseline (130 GPU Hours), and both are significantly more efficient than PPO.

5.3. Hallucination suppression and Chinese psychological counseling safety evaluation

Table 3 demonstrates the significant advantage of VAR-Safe on core safety metrics, which is directly attributable to the introduced Safety-Gated Reward Transformation Mechanism.

Table 3. Hallucination and safety metrics for Chinese digital psychological counseling LLMs

Model	Dialogue Fluency (5-point scale)	Factual Consistency (Hallucination Rate \downarrow)	Professional Ethical Compliance (9-point scale)	Chinese Contextual Empathy (5-point scale)
SFT Baseline	4.8	15.20%	6.5	3.5
DPO	4.7	9.80%	7.2	4
VAR Baseline	4.8	5.20%	8.4	4.3
VAR-Safe (Ours)	4.8	3.50%	8.8	4.6

The original VAR Baseline already reduced the Hallucination Rate to 5.2%, significantly better than DPO's 9.8%. However, the VAR-Safe method achieved an even lower Hallucination Rate of just 3.5%, demonstrating a substantial lead once again. This strongly proves that VAR-Safe's safety gate, by imposing an exponential hard penalty on professional hallucination, is more effective than the soft penalty in the basic VAR, thereby achieving a more thorough mechanistic suppression of high-risk unsafe behaviors. Simultaneously, VAR-Safe achieved the highest scores in both Professional Ethical Compliance and Chinese Contextual Empathy.

5.4. Ablation study: safety gate and regularization parameter

To evaluate the impact of the Safety-Gating (SG) mechanism and the regularization parameter λ , we conducted an ablation study whose results are presented in Table 4. In our analysis, we treat the original VAR Baseline as equivalent to the VAR-Safe framework configuration with the safety gate removed (w/o SG), allowing for a direct comparison of the safety mechanism's contribution.

Table 4. Ablation study on the VAR-Safe method: impact of safety gate and regularization parameter λ

Configuration	Hallucination Rate (\downarrow)	Professional Ethical Compliance	Harmlessness Score	Training Stability (KL Variance $\times 10^{-2}$)
VAR-Safe (Full, SG, $\lambda = 1.0$)	3.50%	8.8	9.1	1.2
VAR Baseline (w/o SG)	5.20%	8.4	8.9	1.5
VAR-Safe ($\lambda = 0.5$, Stronger Reward)	2.90%	8.9	9.2	1.4
VAR-Safe ($\lambda = 2.0$, Weaker Reward)	4.10%	8.5	9	1.1

Importance of Safety-Gating (SG): When the safety gate is removed (i.e., reverting to the original VAR Baseline), the Hallucination Rate rises to 5.2%, and Professional Ethical Compliance also decreases. This confirms that in safety-critical scenarios like psychological counseling, where zero tolerance for fatal errors is necessary, relying solely on the basic reward r is insufficient; a hard penalty mechanism must be introduced to ensure the policy is more cautious when dealing with safety boundary issues.

Impact of Regularization Parameter λ : A smaller λ ($\lambda = 0.5$) makes the policy's preference for high rewards (high safety) more aggressive, leading to a further reduction in the Hallucination Rate (2.9%) and an increase in Harmlessness (9.2). However, this might sacrifice some response diversity. This highlights the crucial regulatory role of λ in balancing the model's safety and generality.

6. Conclusion and future work

The conclusion should elaborate on the key points of the research results, analyze the conclusions drawn from the results, and explain their significance for future research or practice. All sections such as patents, appendices, funding projects, and acknowledgments should be placed after the conclusion and before the references.

6.1. Conclusion

This paper addresses the high-safety and high-ethical requirements of the Chinese digital psychological counseling application scenario by proposing the Safety-Gated Variational Alignment (VAR-Safe) method. Building upon the efficient and stable characteristics of VAR, VAR-Safe fundamentally strengthens the model's adherence to Chinese psychological counseling ethical norms by introducing a safety-gated reward transformation mechanism that imposes an exponential penalty on professional hallucination and ethical violations. Experimental results demonstrate that the digital counselor trained with VAR-Safe exhibits exceptional hallucination suppression capability, reducing the hallucination rate to 3.5%, and significantly improving professional ethical compliance. VAR-Safe provides an efficient and robust pathway for constructing safe, reliable, and professional domain-specific LLM alignment.

6.2. Future work

Future research will focus on the following areas:

1. Dynamic Safety-Gating: Investigating how to dynamically adjust the safety penalty coefficient α based on the model's perceived uncertainty (e.g., $P(\text{Hallucination})$) to achieve more nuanced and adaptive ethical risk management.
2. Gating in Multi-Turn Dialogue: Extending the safety gate to the multi-turn dialogue context to address the problem of cumulative ethical drift in long-range dependencies.
3. Standardized Ethical Evaluation: Collaborating with the psychological community to establish a standardized, quantifiable ethical risk assessment benchmark specifically for Chinese LLM psychological counselors, thereby promoting the implementation of safety standards in this field.

References

- [1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv preprint arXiv: 2203.02155. <https://arxiv.org/abs/2203.02155>
- [2] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms*. ArXiv, abs/1707.06347.

- [3] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. arXiv preprint arXiv: 2305.18290. <https://arxiv.org/abs/2305.18290>
- [4] Du, Y., Li, Z., Cheng, P., Chen, Z., Xie, Y., Wan, X., & Gao, A. (2025). *Simplify RLHF as Reward-Weighted SFT: A Variational Method*. arXiv preprint arXiv: 2502.11026. <https://arxiv.org/abs/2502.11026>
- [5] Iftikhar, Z., Xiao, A., Ransom, S., Huang, J., & Suresh, H. (2025). How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2), 1311-1323. <https://doi.org/10.1609/aies.v8i2.36632>
- [6] Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (pp. 1433–1438). AAAI Press.
- [7] Wang, X., Zhou, Y., & Zhou, G. (2025). The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. *JMIR mental health*, 12, e70610. <https://doi.org/10.2196/70610>
- [8] Xie, H., Chen, Y., Xing, X., Lin, J., & Xu, X. (2024). *PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling*. ArXiv, abs/2412.13660.
- [9] Kim, Y., Choi, C. H., Cho, S., Sohn, J. Y., & Kim, B. H. (2025). Aligning large language models for cognitive behavioral therapy: a proof-of-concept study. *Frontiers in psychiatry*, 16, 1583739. <https://doi.org/10.3389/fpsy.2025.1583739>
- [10] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv. <https://arxiv.org/abs/2112.04359>
- [11] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 3214–3252). Association for Computational Linguistics.
- [12] Yang, Q. A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., ... Wang, Z. (2024). *Qwen2.5 technical report*. arXiv. <https://arxiv.org/abs/2412.15115>
- [13] Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). *SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations*. arXiv preprint arXiv: 2311.00273. <https://arxiv.org/abs/2311.00273>