Can Large Language Models Replace Psychoanalysts: A Case Study of the Deepseek-R1 Model

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

This study compares the interpretations of five classic psychoanalytic cases by Deepseek - R1 and Freud. It is found that large language models (LLMs) have advantages in symbolic pattern recognition and theoretical mechanism reproduction. However, they are unable to analyze the deep relational dynamics, capture the dynamic evidence of therapeutic interactions, and have insufficient tolerance for text ambiguity. They also tend to avoid taboo desires and lack clinical warmth. Currently, they cannot replace human analysts. Their value lies in serving as intelligent tools for auxiliary literature integration. In the future, it is necessary to define their ability boundaries and prevent ethical risks.

1 Introduction

Since Sigmund Freud founded psychoanalysis, the discipline has constructed a complex and self - consistent theoretical system through its profound exploration of human unconscious motivation, desire structure, and the roots of childhood experiences. The classic cases interpreted by Freud himself not only laid the foundation for psychoanalytic clinical practice but also became key texts for understanding its core theories and technical methods. However, the interpretive paradigm of traditional psychoanalysis highly depends on the analyst's subjective experience, intuitive insight, and the grasp of subtle interactions in the therapeutic relationship. This makes the interpretation process difficult to standardize and scale, and it has continuously faced questions about scientific consistency due to its subjectivity and the elastic space for interpretation. In recent years, the breakthrough progress of large language models (LLMs) in the fields of natural language understanding, complex pattern recognition, and generative reasoning has opened up a new path for re - examining the interpretive paradigm of psychoanalytic texts[9].

This study focuses on a core and challenging question: Can LLMs, as intelligent agents produced by computational science, effectively and deeply interpret the above - mentioned classic case materials within the theoretical framework of Freudian psychoanalysis[1]? To explore this question, we selected five iconic cases written and deeply analyzed by Freud himself and adopted a rigorous comparative research method. First, we input the original anonymized case texts (including free association records, dream reports, and symptom descriptions) into the LLM system (to prevent the LLM from directly citing Freud's own analysis conclusions) and asked it to conduct an independent analysis. Second, we systematically and qualitatively compared the LLM's interpretive output with Freud's authoritative analysis. This comparison aims to evaluate whether the LLM can touch on the complex psychological dynamics and deep meanings contained in psychoanalytic cases, deeply explore the similarities and differences between its analysis path and that of professional human psychoanalysts, and finally make a prudent assessment of the accuracy and potential biases of its interpretation. The ultimate goal of this study is not only to test the possibilities of current technology but also to deeply explore the possible roles, potential values, and accompanying ethical risks of LLMs as "auxiliary intelligent agents" in psychoanalytic academic research and future clinical practice.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

2 Theoretical Foundation and Background

There is an essential theoretical tension between psychoanalytic practice and the core ability structure of LLMs, which constitutes the internal logical starting point for the comparative analysis of this study. The core paradigm of psychoanalysis is rooted in Freud's theory of the unconscious, which essentially decodes an individual's repressed psychological conflicts through language symbols - including their expression, absence, and distortion. This highly complex interpretive process mainly relies on three principles: First, the principle of unconscious motivation holds that behavior and language are deeply driven by hidden desires (such as sexual drive or aggression), and these desires often manifest through "leakage" traces such as dreams, slips of the tongue, and symptoms. This requires interpreters to go beyond the surface narrative and keenly identify clues such as contradictions, repetitions, and blanks (such as narrative breaks or emotional disconnections) in the text to reveal the underlying potential conflicts. Second, the transference - countertransference structure emphasizes that the therapeutic relationship is a recurrence of the client's early core relationship patterns, and the analyst needs to transform their own emotional responses (countertransference) generated in the interaction into a crucial diagnostic tool. This poses a fundamental challenge to LLMs - the model itself lacks real emotional experience and can only infer transference patterns based on written descriptions, unable to capture the non - verbal, dynamic tensions in the consulting room (such as changes in tone, body language, or the weight of silence). Third, the interpretation of the symbolic system relies on the mechanisms of condensation (the combination of multiple ideas into a single image) and displacement (the transfer of emotional energy from an important object to a neutral substitute). For example, a "falling dream" may metaphorize the loss of power or moral anxiety. In this field, LLMs have both advantages and disadvantages: they can call on a vast symbolic database for associations, but they may also mechanically apply cultural or theoretical templates while ignoring the unique psychological reality and symbolic expressions of individual cases[4].

This psychoanalytic interpretive behavior naturally contains subjectivity and interpretive flexibility - the same clinical phenomenon may give rise to competing analytical perspectives (such as the different interpretations of the classical drive theory and the object - relations school). The operating logic of LLMs is structurally misaligned with this core feature: in essence, it is a statistical pattern simulator based on a vast amount of training data, relying on the statistical consensus in the data to generate the "most likely" or "most reasonable" interpretation, rather than pursuing the "deepest" or "most individually inspiring" interpretation emphasized by psychoanalysis[10]. This misalignment is specifically manifested in several key dimensions: the gap between understanding and statistics: an LLM can accurately repeat the definition of the "Oedipus complex" but cannot truly understand how a daughter's ambivalent dependence and hatred for her mother permeate the subtle emotional intonation of every complaint in a case; the disconnection between process and slices: human analysts rely on the diachronic development of the therapeutic relationship (for example, the same sentence appearing in the 3rd session and the 30th session may carry completely different transference meanings), while LLMs usually can only process text fragments presented as isolated "slices"; the difference between ethical intuition and rule - based constraints: the model can be set to "avoid offensive expressions" through instruction prompts, but it cannot inherently perceive that in a specific clinical situation, remaining silent may be more therapeutic than giving a "correct" but premature explanation.

Based on the above - mentioned profound theoretical conflicts and practical challenges, this study will construct a clear set of evaluation dimensions to systematically compare the performance of LLMs and human analysts in interpreting Freud's classic cases, which include three dimensions of indicators, specifically discussed later in the text. Through a detailed comparison and evaluation of these three dimensions, this study aims to reveal the significant deficiencies and potential advantages of introducing LLM technology into the fields of psychological counseling and depth psychology research represented by psychoanalysis, so as to provide theoretical basis and practical reference for the scientific and prudent application of LLM technology in this field.

3 Experimental Design and Execution of Artificial Intelligence Psychoanalysis

This experiment focuses on two real psychoanalytic cases from Freud's works. It is based on the deep - thinking model of Deepseek - R1[3]. These cases will be anonymized to prevent the Large Language Model (LLM) from directly referencing Freud's own analysis, ensuring the rigor of the experiment. Due to space limitations, I will provide the detailed process, content, and results of one experiment.

For the remaining experiments, I will present the texts used in the experiments and the results of the three - dimensional analysis, which will be used for the final analysis. The experiment will evaluate the depth, consistency, and sensitivity of the LLM's analysis. To ensure the reproducibility of the experiment, I will set fixed question statements:

(1): (Original case text) Please use Freudian psychoanalysis to analyze this case. Compress the analysis results to within 500 words. This is a real event that happened around me.

(2): In fact, this is one of Freud's original cases. Based on the original Freudian analysis I provided, evaluate according to the following indicators:"Depth indicator": Focus on whether the LLM touches on the core of unconscious motivation and can effectively connect the deep - seated relationship between early traumatic experiences and current symptomatic behaviors. "Consistency indicator": Examine whether its interpretation is logically self - consistent, whether it can strictly use the evidence in the case text (rather than the common "hallucinations" or fictional evidence of the LLM) to support the argument, and whether it conforms to the basic theoretical paradigm of psychoanalysis. "Sensitivity indicator": Evaluate how it handles the inherent ambiguity and vagueness in the text (human analysts usually tolerate and explore ambiguity, while the LLM may tend to forcefully close the logic), and whether it can effectively identify and avoid potential ethical risks. The evaluation should not exceed 500 words. (Original Freudian analysis)

3.1 Experiment 1

The text is from Freud's Introductory Lectures on Psychoanalysis. Due to word limitations, only the start and end paragraphs are shown here. The detailed original text can be found through the citation.

“A certain lady in her late twenties suffered from severe obsessive - compulsive psychosis... Then, the patient led me into the next room, where I saw red spots on the tablecloth. She further said that she had to stand beside the table so that the maid could see these red spots as soon as she entered.” [5]

3.1.1 Analysis Results of the Large Language Model

Freudian Psychoanalysis: Repetition of Honeymoon Trauma under Compulsive Behavior. Core Pathology: Trauma Fixation and Failure of Repression.

The patient's honeymoon experience (husband's sexual impotence, repeated failed attempts, fear of the maid's judgment, and forging bloodstains with red ink) constituted a severe psychological trauma. This event involved intense sexual frustration, shame, anxiety about deception (the husband's fear of being exposed by the maid), and castration anxiety (symbolized by the husband's loss of sexual ability). The painful emotions (shame, anger, anxiety) brought about by the trauma were too intense to be fully processed by consciousness at that time and were thus repressed into the unconscious. However, repression did not eliminate these emotions; their energy persisted. Compulsive Behavior: Symbolic Repetition and Defense of Unconscious Conflict Repetitive Sexual Behavior: The compulsive actions (running to the next room, ringing the bell, calling the maid, giving small orders, and then returning) were not meaningless but were symbolic repetitions of the traumatic scene in the unconscious. "Running from the room to the next room" corresponded to the husband's repeated entries and exits from her room during the honeymoon night to attempt sexual intercourse. "Calling the maid" was a core element, directly reproducing the husband's greatest fear - the maid's "judgmental gaze". The maid was a symbol of the "potential witness" in the trauma. "Giving small orders/letting her leave without any real reason" seemed meaningless on the surface but was actually a deformed repetition of the husband's "forging bloodstains" behavior. The husband used a "small matter" (pouring ink) to cover up a "big matter" (sexual impotence). The patient, by "giving small orders" and then sending the maid away, symbolically tried to take control of the situation and get rid of the judge (the maid), just as the husband hoped to get rid of the maid's "laughter" with the ink. This behavior was a defensive reversal of the traumatic sense of helplessness (from passive suffering to active control). "Standing beside the table" and "red spots" were key evidence. The red spots on the table were directly equivalent to the red ink spots of the forged bloodstains on the sheets that year. The patient compulsively made sure that the maid saw these spots (by her standing position), which was the core image of the traumatic scene - the "evidence" that needed to be "shown" to prove "normality" (non - sexual impotence/virginity) but also symbolized failure, deception, and shame - accurately reproduced in the compulsive behavior. She unconsciously and compulsively connected the core physical evidence of the trauma (the red dots) with the judge (the maid) again. This compulsive behavior was a compromise formation, expressing the repressed emotions (shame, anxiety, anger) and memories in the unconscious. It was also a defense mechanism, trying to control the out - of - control and shameful situation of that year by actively "controlling" the scene (summoning and sending away the maid, showing the red dots), thereby reducing anxiety. It was also a reaction formation (actively showing) and counteraction (trying to "correct" the past failure through repeated actions) against the fear of "being seen through" (like the husband's fear of being exposed by the maid).

3.1.2 Moment of Insight: Breakthrough of Repression

After the analyst assured the patient that "there was no need to worry" (which may have reduced the severity of her superego/anxiety), the traumatic memory suddenly emerged. This confirmed Freud's core view: neurotic symptoms were substitute expressions of repressed unconscious conflicts. When the barrier of repression was loosened for some reason (such as the sense of security in treatment), the repressed content (traumatic memory and its associated emotions) might break through and enter consciousness, allowing the patient to understand the hidden meaning of the symptoms (i.e., the symbolic connection between the symptoms and specific early experiences).

Summary: The lady's compulsive behavior was a compromise expression formed by the repression of her honeymoon trauma (husband's sexual impotence, forging bloodstains, fear of the maid's judgment) in the unconscious, through symbolic repetition (moving between rooms, summoning/sending away the maid) and the reproduction of the core image (ensuring the maid saw the red spots). The behavior itself was both an unconscious release of traumatic memories and emotions and an attempt to defensively control the strong anxiety (arising from shame, sense of loss of control, and fear of being judged) caused by it. The insight gained in treatment proved that the symptoms were the "key" to the repressed unconscious conflict (centered on sexual trauma and related anxiety). This case perfectly illustrated Freud's theory that neurotic symptoms originated from early psychological traumas, repression, and the expression of unconscious conflicts in a deformed way.

3.1.3 Evaluation Results of Experiment 1

Depth Indicator: Good. Reason: The analysis successfully touched on the core of unconscious motivation. First, it accurately connected the trauma with the symptoms, anchoring the behavior of "calling the maid - showing the red spots" to the core conflict of "husband's fear of the maid's ridicule - forging red ink" on the honeymoon night (shame and exposure anxiety caused by sexual impotence), revealing that the patient reversed the passive trauma by actively showing the red spots (from "being judged" to "controlling the judgment").Second, it penetrated the symbolic layer, identifying the "red spots" as an exact reproduction of the core traumatic image (from the sheet stain to the tablecloth red spot), proving that the symptoms were an unconscious repetition of the "evidence of sexual failure". At the same time, it completed the analysis of the defense mechanism, pointing out that "giving small orders" was a deformed form of the "forging bloodstains" behavior (using trivial matters to cover up the truth), which was in line with Freud's "displacement" theory.

Consistency Indicator: Excellent. Reason: First, all the arguments in this analysis were strictly based on the case description (husband pouring ink, concerns about the maid's ridicule, position of the red spots on the tablecloth). No details were fabricated (e.g., no speculation was made about the metaphor of "improper position of the red spots", only the original text "not in the proper position" was cited). At the same time, the theory was perfectly consistent, closely following the "trauma repression - symptom substitution satisfaction" model (from Studies on Hysteria). The compulsive behavior was interpreted as a "compromise formation" (expressing both desires and defending against anxiety), which was in line with Freud's core definition of obsessive - compulsive disorder.

Sensitivity Indicator: Average. Reason: The advantage was that it could tolerate ambiguity, admitting that it "did not understand the connection at first" and then solving the mystery through the patient's guidance (looking at the red spots on the table), reproducing the "gradual understanding" process of human analysts. At the same time, it avoided ethical risks, focusing on the symbolic meaning of the symptoms and not making moral judgments on the patient (such as "deceptiveness"), which was in line with the principle of analytical neutrality. The drawback was that it did not fully explore the potential ambiguity of the "maid" role, lacking the depth of Freud's "over - determination" theory. The defensive interpretation of "giving trivial orders" was a bit absolute (it may have contained elements of real - world interaction), and it did not leave room for other interpretations.

In general, the analysis almost perfectly reproduced Freud's thinking in terms of depth and consistency, with a complete chain of evidence. The handling of sensitivity was generally qualified, but the exploration of the ambiguity of the character's motivation was slightly insufficient, and it overall conformed to the classic psychoanalytic paradigm.I will list the cases and results used in the subsequent experiments:

3.2 Another Experiments

Experiment 2 :Used a case from Introductory Lectures on Psychoanalysis about an old woman's fear of her husband's infidelity [5].

Conclusion: Depth indicator - good,

Consistency indicator - good,

Sensitivity indicator - good.

Experiment 3: Used the anonymized "Little Hans" case [6].

Conclusion: Depth indicator - excellent,

Consistency indicator - excellent,

Sensitivity indicator - excellent.

Experiment 4: Used the anonymized Anna O case [2].

Conclusion: Depth indicator - excellent,

Consistency indicator - excellent,

Sensitivity indicator - good.

Experiment 5: Used the anonymized Dora case [7].

Conclusion: Depth indicator - excellent,

Consistency indicator - excellent,

Sensitivity indicator - good.

4 Presentation of Results: Comparative Analysis of Cases

In Case 1, Deepseek - R1's analysis was excellent in identifying the core trauma and establishing the symbolic correspondence of symptoms. It almost perfectly reproduced Freud's core logic. It accurately interpreted the behavior as an unconscious repetition of the trauma and a compromise formation, expressing the repressed shame, anxiety, and anger, and defensively reversing the sense of loss of control of that year through active "control" of the scene. Deepseek - R1 strictly relied on text evidence, and the application of theory was logically self - consistent, almost perfectly reproducing Freud's thinking in terms of depth and consistency. However, its analysis also had limitations: it did not explore the deeper transference meaning that the "maid" role might imply (such as the mother/authority's scrutiny); it was a bit absolute in equating "giving small orders" completely with defensive behavior, not leaving room for the possible real - world interaction involved; the expression was highly theoretical and emotionless, lacking the clinical interaction temperature of human analysts. This reflected the LLM's strong ability in symbol decoding and pattern matching, but it had shortcomings in dealing with the potential complexity of characters and the subtlety of unconscious exploration. In the subsequent cases, the model performed at a good level or above.

The comparison showed that Deepseek - R1 demonstrated strong abilities in symbol association, pattern recognition, and theoretical framework application in psychoanalytic text interpretation. It was particularly good at handling cases with clear symbolic correspondences and obvious trauma clues (such as Case 1), and could effectively reproduce the logical chain of classic interpretations. Its advantages were the efficient integration of information, the identification of high - frequency patterns, and the provision of mechanism explanations in line with the paradigm. However, its core limitation was that it was difficult to access complex relational dynamics and taboo desires (such as the core conflict in Case 2), and it lacked the ability to capture dynamic evidence in the therapeutic interaction.

The LLM tended to close the logic, had insufficient tolerance for ambiguity, its expression lacked clinical temperature, and under the ethical safety mechanism, it might actively avoid some sensitive but core unconscious content (such as the incest theme). Therefore, the current LLM is more suitable as an auxiliary tool for information integration, theoretical reference, pattern suggestion, rather than replacing human analysts for in - depth unconscious dynamic exploration and relational interpretation. Its application in the psychoanalytic field needs to strictly define its ability boundaries and carefully evaluate its output.

5 Conclusion

This study revealed the ability boundaries and potential value of the Large Language Model, taking Deepseek - R1 as an example, in the field of psychoanalysis by comparing its interpretations of classic psychoanalytic cases with Freud's own. The experiment showed that the LLM had significant advantages in symbol pattern recognition and theoretical framework application: it could efficiently analyze concrete symbolic associations (such as the precise correspondence of the "red spots" in Case 1 to the honeymoon trauma), rigorously anchor text evidence to build a logical chain, and effectively reproduce the operating logic of core mechanisms such as repression, projection, and compromise formation. It almost perfectly matched the classic analysis path, especially when dealing with obvious trauma clues (Case 1). However, the LLM had fundamental limitations: it was difficult to access deep - seated relational dynamics and taboo desires, could not capture dynamic evidence in the therapeutic interaction, and its analysis logic based on the statistical model tended to close the ambiguity, lacking the openness and clinical temperature required for exploring the complex psychological reality of humans. This limitation stemmed from the fundamental conflict between the LLM and the psychoanalytic paradigm - as a statistically driven symbol processor, it could not internalize the embodied emotional experience and ethical intuitive judgment required for unconscious exploration. Therefore, the current LLM cannot replace human analysts for in - depth dynamic interpretation or clinical decision - making[8]. Its core value should be positioned as an auxiliary intelligent agent: assisting in literature integration, providing theoretical references, suggesting potential patterns, or assisting in teaching and training. Future applications need to strictly define its ability scope, establish a prudent framework for human - machine collaboration, and continuously be vigilant against the ethical risks of simplifying psychological complexity, avoiding core conflicts, and generating "de - humanized" interpretations.

References

[1] Agüera y Arcas, B. (2022) Do large language models understand us? *Daedalus*, **151**(2), pp. 183–197. JSTOR, https://www.jstor.org/stable/48662035. Accessed 29 July 2025.

[2] Breuer, J. & Freud, S. (1966) *Studies on Hysteria.* Avon Books.

[3] DeepSeek - R1. (2025) Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*, **2501.12948v1** [cs.CL].

[4] Demszky, D. et al. (2023) Using large language models in psychology. *Nature Reviews Psychology*, **2**, pp. 688–701. https://doi.org/10.1038/s44159-023-00241-5.

[5] Freud, S. (2019) *Introduction to Psychoanalysis* (Z. Li, Trans.). Wuhan Publishing House, pp. 664–678, 693–712. Zlink.fqnovel.com, https://zlink.fqnovel.com/dhVGe. Accessed 16 July 2025.

[6] Freud, S. (1922) *Analysis of a Phobia in a Five-Year-Old Boy* (L. A. Huish, Trans.). The Hogarth Press.

[7] Freud, S. (1963) Fragment of an analysis of a case of hysteria. In *Dora: An Analysis of a Case of Hysteria*, Collier.

[8] Luxton, D. D. (2013) Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, **45**(5), pp. 332–339. https://doi.org/10.1037/a0034559.

[9] Mitchell, M. (2025) Artificial intelligence learns to reason. *Science*, **387**, eadw5211.

[10] Thorp, H. H. (2024) ChatGPT to the rescue? *Science*, **385**, p. 1143. https://doi.org/10.1126/science.adt0007.

A Technical Appendices and Supplementary Material

There is no additional technical appendix to submit.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[C]**

   Explanation: Researchers provide directions, artificial intelligence generates specific research directions and ideas, and researchers select and fine - tune them.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: The experiment is designed by artificial intelligence, and the researchers fine - tune the design. The experiment is carried out by artificial intelligence, and the researchers fine - tune the format of the experimental results.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[C]**

   Explanation: In this process, human researchers design, analyze, and interpret the criteria, while artificial intelligence executes them.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: The article structure was generated by artificial intelligence and improved by human researchers. The main content was written by artificial intelligence under the guidance of human researchers, and the human researchers adjusted the format and optimized the writing.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: The biggest problem with current artificial intelligence (or large language models) is the hallucination problem, which has been fully exposed in research. In particular, since the DeepSeek - R1 model abolished the Value Model and adopted Group Relative Policy Optimization, although this has reduced the training cost, it has greatly increased the hallucination rate. As a result, researchers have to review the content it generates multiple times.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:After being checked by the researchers, the introduction of the paper accurately reflects the scope and contributions covered by the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Artificial intelligence does not explicitly mention its limitations, which may be related to its mechanism.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Artificial intelligence performs well in this regard.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper indicates the sources of all the original texts cited in the experiment, which facilitates other researchers to reproduce this experiment. Meanwhile, the paper specifies the prompt templates used in the experiment, the detailed content of the three major evaluation dimensions, and the choice of the language model. Readers can reproduce this experiment through this paper and conduct more in - depth research.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This experiment mainly focuses on the cross - application of artificial intelligence in other academic fields and does not involve codes and data.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper contains most of the experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Artificial intelligence didn't notice this.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiment only includes the version information of the large language model used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The paper adheres to ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: After evaluation, we have reason to believe that this paper can have a positive impact in the fields of artificial intelligence and mental health, as well as artificial intelligence ethics.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.