

PsychoLex: Unveiling the Psychological Mind of Large Language Models

Mohammad Amin Abbasi¹, Farnaz Sadat Mirnezami², Hassan Naderi¹

¹Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

²Department of Computer Engineering, University of Guilan, Rasht, Iran

m_abbasi1378@comp.iust.ac.ir

farnaz.mirnezami@gmail.com

naderi@iust.ac.ir

Abstract

This paper explores the intersection of psychology and artificial intelligence through the development and evaluation of specialized Large Language Models (LLMs). We introduce PsychoLex¹, a suite of resources designed to enhance LLMs' proficiency in psychological tasks in both Persian and English. Key contributions include the PsychoLexQA dataset for instructional content and the PsychoLexEval dataset for rigorous evaluation of LLMs in complex psychological scenarios. Additionally, we present the PsychoLexLLaMA model, optimized specifically for psychological applications, demonstrating superior performance compared to general-purpose models. The findings underscore the potential of tailored LLMs for advancing psychological research and applications, while also highlighting areas for further refinement. This research offers a foundational step towards integrating LLMs into specialized psychological domains, with implications for future advancements in AI-driven psychological practice.

1 Introduction

The rise of Large Language Models (LLMs) has significantly advanced artificial intelligence (AI), providing remarkable capabilities in natural language processing and understanding (Guo et al., 2023; Minaee et al., 2024; Wu et al., 2023). These models have shown proficiency in generating human-like text, translating languages, and engaging

in sophisticated dialogues (Agrawal, 2023). However, as users increasingly rely on LLMs for psychological and therapeutic questions (Lai et al., 2023), the limitations of these models in specialized domains have become apparent. Notably, there is a critical absence of datasets designed to evaluate and enhance LLMs' performance in the field of psychology.

Despite considerable progress in general AI research, the integration of psychological expertise into LLMs remains underdeveloped. Existing methodologies often lack the depth required to understand and respond accurately to complex psychological inquiries. Moreover, the field is hindered by the lack of comprehensive datasets that include not only questions and answers but also instructional content tailored to psychological contexts. This gap is significant because it restricts the practical applications of LLMs in psychological research, therapy, and education, where nuanced and precise information is essential.

Our research seeks to address this gap by introducing PsychoLex, a suite of resources and models specifically designed for psychological applications in both Persian and English. The primary objectives of this study are to develop and evaluate specialized datasets, namely PsychoLexQA and PsychoLexEval, and to introduce PsychoLexLLaMA, an LLM developed for psychological tasks. These contributions include: (i) PsychoLexQA, which provides comprehensive instructional content and detailed questions and answers to enhance LLM training; (ii) PsychoLexEval, a multiple-choice question and answer (MCQA) dataset designed for rigorous evaluation

¹<https://huggingface.co/collections/aminabbasi/psycholex-66b64e3768da519596e49de9>

of LLMs in psychological contexts, ensuring they can handle complex psychological queries accurately and contextually;(iii) PsychoLexLLaMA, which improves the performance of LLMs in psychological tasks through continual pre-training and fine-tuning of LLaMA 3.1.(Dubey et al., 2024) Together, these contributions aim to provide robust solutions to existing challenges, enhancing the accuracy and relevance of AI-driven psychological tools and paving the way for future advancements in integrating AI with psychological practice.

The structure of this paper is organized as follows: Section 2 reviews related work in LLMs and their applications in psychology. Section 3 details the datasets developed for this study, including their creation and validation processes. Section 4 discusses the development and fine-tuning of the PsychoLexLLaMA model. Section 5 presents the evaluation methodology and results, comparing PsychoLexLLaMA with other state-of-the-art models. Section 6 provides a comprehensive discussion of the findings, and Section 7 concludes the paper with insights into future research directions and potential applications.

By exploring the intersection of AI and psychology, this paper aims to unveil the psychological capabilities of LLMs and demonstrate their potential to advance both fields significantly.

2 Related Works

In this section, we review existing research that benchmarks the capabilities of large language models (LLMs) in Persian, followed by studies that explore the integration of LLMs into psychological research and applications. This dual-focus review establishes the context for our work, emphasizing both the linguistic challenges specific to Persian and the broader implications of applying LLMs in the field of psychology.

2.1 Benchmarking Large Language Models for Persian

Recent advancements in large language models (LLMs), particularly ChatGPT, have generated significant interest in their evaluation across various languages and tasks. ChatGPT's performance on various Persian natural language processing tasks is evaluated by Abaskohi et al. (2024). they present a comprehensive evaluation of large language models (LLMs) for the Persian language, focusing on models like GPT-3.5-turbo (OpenAI,

2023a), GPT-4 (OpenAI, 2023b), and OpenChat-3.5. This study, which is the first extensive benchmarking effort for Persian, aims to address the challenges posed by Persian as a low-resource language with unique linguistic features. The evaluation covers a broad range of natural language processing (NLP) tasks, including sentiment analysis, question answering, natural language inference, and translation. the study highlights the model's superior performance in multiple-choice questions(MCQs) related to math and general knowledge from the ParsiNLU dataset (Khashabi et al., 2020). These benchmarks are particularly important for assessing the models' reasoning capabilities in Persian. While ChatGPT-4 excels across several benchmarks, its application in psychology has not been tested, underscoring a critical area for future research.

2.2 Khayyam Challenge (PersianMMLU)

Recent advancements have focused on optimizing the performance of Large Language Models (LLMs). The PersianMMLU (Ghahroodi et al., 2024) is particularly significant as it concentrates on the Persian language capabilities of these models. It evaluates their proficiency in answering multiple-choice questions across diverse fields such as mathematics, science, logic, and intelligence testing. This comprehensive evaluation involved advanced models like GPT-3.5, GPT-4(OpenAI, 2023b), Aya (Ustun et al., 2024), PersianMind(Rostami et al., 2024), mT0 (Muennighoff et al., 2023), mGPT (Shliazhko et al., 2022), and Claude3-haiku (Anthropic, 2024). The study utilized a robust dataset derived from Iran's national university entrance exams and educational assessments. While GPT-4 emerged as the superior model, its efficacy in psychological applications remains untested. This gap highlights the necessity of our current research, which aims to specifically evaluate the performance of LLMs in psychology-related scenarios.

2.3 Using large language models in psychology

Demszky et al. (2023) explores the integration of LLMs, particularly GPT-3 and GPT-4, into psychological research practices. These models' adeptness at text generation, dialogue engagement, persona simulation, and information synthesis provides innovative approaches to studying various psychological subfields. The primary aim

is to evaluate the extent to which LLMs can enrich psychological research methodologies. Despite their potential, LLMs often fall short in delivering contextually accurate advice consistently. This study highlights the importance of refining LLMs through fine-tuning and reinforcement learning from human feedback to ensure their practical efficacy in real-world psychological settings. The extensive datasets used to train these models, encompassing diverse sources of human language data, are aimed at tailoring LLMs to better serve both theoretical and applied psychology.

2.4 Exploring the Frontiers of LLMs in Psychological Applications

The application of Artificial Intelligence (AI), especially large language models (LLMs), is revolutionizing psychological research. A study by [Ke et al. \(2024\)](#) underscores significant advances in language models and their profound impact on the field of psychology. LLMs like OpenAI's ChatGPT facilitate various research activities, including literature reviews, hypothesis formulation, experiment design, data analysis, and scholarly writing across several psychological domains such as cognitive, behavioral, clinical, educational, developmental, and social psychology. While these models offer substantial benefits, the review also delineates key technical and ethical challenges, including data privacy concerns and inherent limitations of LLMs. The authors advocate for the careful integration of these technologies in psychological research to enhance our understanding of the human mind and improve the methodologies employed in psychological studies.

In conclusion, the existing body of research on large language models (LLMs) highlights both significant advancements and notable limitations in applying these models to Persian natural language processing and psychological applications. However, challenges persist due to the low-resource nature of the Persian language and the limited evaluation of LLMs in psychological settings. While LLMs hold potential for improving psychological research, they often lack consistency and contextual accuracy. Future work is needed to refine these models for specialized fields, addressing their current limitations and improving their reliability in practical applications.

3 Dataset

This section outlines the datasets developed to investigate the application of large language models (LLMs) in psychology. We detail the creation and utilization of three pivotal datasets: the foundational pretraining data, the PsychoLexQA dataset for instructional content, and the PsychoLexEval dataset for evaluating model comprehension and performance.

3.1 Pretraining Data

For the pretraining phase, we employed "Introduction to Psychology" by Hilgard (1953), a seminal textbook noted for its comprehensive insights into psychology. This text was used in both its Persian and English versions to establish a bilingual foundation for our models. The dataset comprised approximately 1.3 million tokens, offering a rich and diverse corpus that spans a broad spectrum of psychological topics. This extensive pretraining data enabled our models to develop a deep understanding of essential psychological concepts and terminology, facilitating their application in both Persian and English contexts.

3.2 PsychoLexQA

For the instructional dataset, we adopted two distinct methodologies to generate detailed and comprehensive instructional content in both Persian and English

3.2.1 Document-Based Instructions

The first method involved extracting instructional content from "Introduction to Psychology" in both languages. This process was automated using the GPT-4o model, where paragraphs from the textbook were analyzed to grasp key concepts. For each paragraph, the model generated a series of questions and answers aimed at testing material comprehension. Each question was crafted to be clear and precise, with detailed answers provided to ensure a thorough understanding of the discussed psychological concepts. Paragraphs lacking sufficient content for question generation were identified and noted. This method resulted in a dataset containing 7,055 entries. Appendix A demonstrates an example of PsychoLexQA dataset based on psychology document.

3.2.2 Self-Instruct

The second method focuses on creating structured instructional tasks for various psychological subcategories in both Persian and English. This semi-automated approach, inspired by the Self-Instruct framework (Wang et al., 2022), utilizes a predefined list of subcategories such as Clinical and Cognitive Psychology and tasks such as Case Study Analysis and Experiment Design to generate detailed bilingual task instructions. Through iterative processes, the model generates comprehensive inputs and outputs, resulting in a dataset of 3,001 entries covering multiple psychological domains, ensuring a diverse and realistic range of scenarios. Figures 3, 4, 5, 6 in the Appendix A depicts several examples of PsychoLexQA dataset which has been created using a self-instruct method.

3.2.3 Scope of PsychoLexQA

The PsychoLexQA dataset, focuses on simulating real-world psycho-logical scenarios through a range of instructional tasks. These tasks challenge LLMs with exercises like: (1) Case Study Analysis: Identifying psychological issues and proposing interventions; (2) Experiment Design: Developing experimental methodologies, hypotheses, and variables; (3) Data Interpretation: Analyzing statistical or experimental data; (4) Scenario Response: Offering interventions for hypothetical psychological scenarios; (5) Theoretical Explanation: Explaining psycho-logical theories or concepts; (6) Diagnostic Evaluation: Simulating diagnostic processes for psychological conditions; (7) Ethical Dilemma Resolution: Proposing ethical resolutions in psychological practice; (8) Intervention Planning: Crafting treatment plans for psychological conditions; (9) Literature Review: Summarizing key findings from psychological research; (10) Comparative Analysis: Comparing psychological theories, interventions, or case outcomes.

The dataset further breaks down into subcategories that reflect distinct psychological areas, such as Clinical Psychology, Cognitive Psychology, Developmental Psychology, Social Psychology, Biological Psychology, Industrial-Organizational Psychology, Neuropsychology, Health Psychology, and Educational Psychology. Each subcategory is paired with tasks de-signed to simulate challenges and interventions relevant to that field. Together, both datasets provide a comprehensive tool for

evaluating and training LLMs in diverse and realistic psycho-logical scenarios, with PsychoLexEval assessing broad understanding and PsychoLexQA focusing on task-based problem-solving.

3.3 PsychoLexEval

The PsychoLexEval dataset, a multiple-choice question and answer (MCQA) format in both Persian and English, is designed to assess the comprehension and performance of LLMs in psychology. This section will describe the data collection and review process, the methods employed to ensure quality and compliance, and the broad scope and coverage of this MCQA dataset. An example of the PsychoLexEval dataset can be found in Appendix A.

3.3.1 Data Collection

To construct this dataset, we compiled questions from multiple significant sources: (1) Graduate Entrance Exams: questions from psychology entrance exams (2014-2024) that cover advanced topics; (2) Employment Exams: questions from various job tests, including both specialized and general psychology; (3) Online Sources: Questions from trusted psychology test websites; (4) GPT-4 Generated Content: questions from Psychology books, covering a wide range of topics.

3.3.2 Filtering and Review

To ensure high quality and legal compliance, we implemented rigorous filtering and review processes for the dataset. Initially, a human review was conducted where a sample of questions was meticulously scrutinized by experts. This step was crucial to ensure that each question was relevant, complete, and clearly articulated. During this phase, we specifically retained only those questions that had exactly four answer options, ensuring consistency and clarity in the evaluation process. Additionally, to avoid any legal complications, we carefully removed any content that potentially violated copyright laws. This step was essential to maintain the integrity of the dataset and ensure that all included materials were legally compliant for use in our research and broader academic dissemination. These measures collectively reinforced the dataset's reliability and adherence to legal standards, providing a robust foundation for evaluating large language models within psychological contexts.

3.3.3 Scope of PsychoLexEval

The PsychoLexEval dataset is meticulously designed to evaluate the comprehension and performance of large language models (LLMs) in psychology, encompassing a broad spectrum of psychological fields. These include general psychology, developmental psychology, and clinical psychology, along with psychometrics, cognitive tests, and industrial-organizational psychology. The dataset also covers social and educational psychology, exceptional children’s needs, and foundational concepts from Introduction to Psychology. With 3,430 rows of diverse content, it serves as an invaluable resource for researchers to assess LLMs’ capabilities across various psychology domains.

4 PsychoLexLLaMA

In this section, we detail the development of PsychoLexLLaMA, a specialized large language model (LLM) designed explicitly for psychology. Our goal was to surpass the performance of general-purpose models by optimizing our model to require minimal data and hardware resources. The model benefits significantly from the integration of the PsychoLexQA dataset, which includes tasks such as case study analysis, experiment design, and diagnostic evaluation. This dataset allows PsychoLexLLaMA to handle complex psychological scenarios across a variety of domains like clinical, cognitive, and developmental psychology, making it a highly valuable tool for both practical and research-oriented applications.

The integration of PsychoLexQA significantly enhances PsychoLexLLaMA’s role in applied psychology and psychological science. It assists clinical practitioners with diagnostics and interventions, supports research with experiment design and data interpretation, and enriches education by generating tailored instructional content. Overall, it provides specialized, accurate responses, making it a valuable tool across clinical, research, and educational domains in psychology.

We utilized the Transformers² library for model development. The process of constructing our model is illustrated in Figure 1.

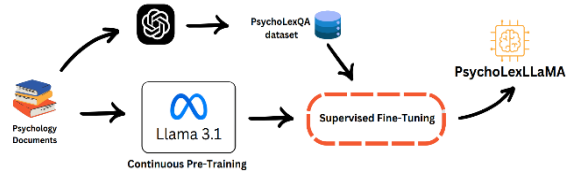


Figure 1: Process of constructing PsychoLexLLaMA model.

4.1 Continuous Pre-Training

For continuous pre-training (Zhou et al., 2024), we employed the LoRA technique (Hu et al., 2021) on the bilingual texts of "Introduction to Psychology" by Hilgard. This foundational work was processed in both Persian and English, leveraging the established pretraining data. We utilized LLaMA 3.1 (Dubey et al., 2024) as our base models in two configurations: 8B and 70B. This stage was critical for aligning the base models with psychological content, thereby enhancing their understanding and application of complex psychological concepts efficiently. The pre-training for the 8B model took 8 minutes using a single A100 80GB GPU, while the 70B model required 41 minutes on two A100 80GB GPUs. Table 1 provides a detailed overview of the LoRA training configurations used during this phase.

Lr	Rank	Alpha	Dropout
1e-5	8	16	0.0

Table 1: LoRA training configurations

4.2 Supervised Fine-Tuning

The supervised fine-tuning phase was essential for tailoring our models to meet the specific demands of psychological analysis. Utilizing the PsychoLexQA dataset, which includes both instructional content and a comprehensive set of questions and answers, we applied the LoRA technique to further train the pre-trained models. This phase was pivotal in refining the models’ abilities to interpret and respond accurately to intricate psychological queries and scenarios within the dataset. The supervised fine-tuning for the 8B model took 22 minutes using a single A100 GPU, while the 70B model required 32 minutes on two A100 GPUs. The LoRA training configurations

² <https://github.com/huggingface/transformers>

used during this phase were the same as those in the continuous pre-training.

4.3 Linear Weight Combination

To bolster the final model’s robustness and preserve the integrity of previous training advances, we implemented a linear weight combination strategy. This involved merging the weights of the LLaMA 3.1 Instruct model with our continuously pre-trained and finely-tuned models. Each model contributed 50% of its weight to the final composite. This method synergistically combined the foundational capabilities of LLaMA with our newly developed psychological expertise, producing a balanced and potent tool adept at handling sophisticated psychological inquiries.

Through these meticulous steps, PsychoLexLLaMA has been meticulously tailored to meet the unique needs of psychological applications. It stands as a robust resource for researchers and practitioners in both psychology and artificial intelligence, providing a reliable platform for further explorations and advancements in these fields. The next sections will evaluate PsychoLexLLaMA’s performance in detail, comparing it with other models to underscore its enhanced capabilities in the realm of psychological research and practice.

5 Evaluation

In this study, we conducted a comprehensive evaluation of various language models that operate in both Persian and English, focusing on their ability to understand and accurately respond to psychological questions. The models assessed include include Qwen2 (Yang et al., 2024), Aya-23 (Aryabumi et al., 2024), Phi-3 (Abdin et al., 2024), Llama-3, Llama-3.1 (Dubey et al., 2024), Gemma 1.1 (Team et al., 2024), command-r, PersianLLaMA (Abbasi et al., 2023), PersianMind (Rostami et al., 2024b), and PsychoLexLLaMA. Our focus on open-source models was intended to enhance the accessibility and reproducibility of our findings. The generation configuration for all the LLMs evaluated is consistent across the experiments and is detailed in Table 2.

5.1 Zero-shot Setting

In the zero-shot setting, models were tested without any prior contextual examples, relying solely

Temp	Max new tokens	top p	Do sample
0.01	16	0.9	True

Table 2: Generation configurations for all evaluated LLMs.

on their pre-existing knowledge. This setting evaluated the models' intrinsic ability to generate accurate responses based solely on their training.

5.2 One-shot Setting

The one-shot setting involved presenting each model with a single relevant example before it answered a question. This setting was used to assess the impact of a minimal context on the accuracy of the models, providing insights into their ability to leverage new information quickly

5.3 Five-shot Setting

In the five-shot setting, models were given five related examples before responding to questions. This scenario tested the models' capacity to utilize more extensive contextual information to enhance their accuracy, offering a deeper understanding of their learning capabilities.

5.4 Evaluation Metric

The effectiveness of each model across the zero-shot, one-shot, and five-shot settings was measured using accuracy as the primary metric. Accuracy was defined as the proportion of correct answers provided by the models relative to the total number of questions posed. This rigorous evaluation approach allowed us to discern the strengths and weaknesses of each model in processing and understanding psychological content comprehensively.

Through these methodical evaluations, we aimed to illustrate the varying capabilities of each model under different contextual conditions. This analysis not only sheds light on how models adapt to incremental information but also highlights their potential applicability in psychological settings, where understanding nuanced human behavior is crucial.

6 Results

This section outlines the outcomes of our evaluation of selected large language models (LLMs) using the PsychoLexEval dataset in both Persian and

English. The primary focus was on assessing the models' proficiency in understanding and responding to psychological questions.

Tables 3 and 4 illustrate the accuracy results of the models on the PsychoLexEval dataset for Persian and English, respectively. These tables quantify how effectively each model comprehends and addresses psychology-related questions across languages.

6.1 Discussion

The results from Tables 3 and 4 provide significant insights into the performance of various LLMs, showcasing their competencies in both Persian and English. Notably, these findings highlight the influence of model architecture and parameter size on handling specialized tasks, such as interpreting and responding to psychology-related questions.

Model	# Param	Accuracy			
		0-shot	1-shot	5-shot	Avg
Qwen2 Instruct	7B	03.55	06.18	08.63	6.12
Gemma 1.1 it	7B	43.07	40.68	27.57	37.11
PersianMind	7B	35.78	35.96	24.63	32.12
Aya-23	8B	39.64	41.42	27.02	36.03
Llama-3 Instruct	8B	33.88	10.66	34.49	26.34
Llama-3.1 Instruct	8B	45.89	41.36	35.78	41.01
PsychoLexLLaMA-pretrain-sft	8B	47.30	43.13	46.61	45.68
PsychoLexLLaMA-average	8B	48.52	41.97	47.05	45.85
PersianLLaMA	13B	20.13	18.52	19.89	19.51
Aya-23	35B	21.07	10.47	22.69	18.08
c4ai-command-r-v01	35B	35.96	21.75	46.20	34.64
Llama-3 Instruct	70B	19.54	09.31	0.5	9.78
Llama-3.1 Instruct	70B	70.34	67.83	70.40	69.52
PsychoLexLLaMA-pretrain-sft	70B	67.79	45.34	68.07	60.4
PsychoLexLLaMA-average	70B	65.84	53.06	69.66	62.85
Qwen2 Instruct	72B	31.37	05.82	50.3	29.16

Table 3 : Accuracy of LLMs on the PsychoLexEval dataset in Persian.

Model	# Param	Accuracy			
		0-shot	1-shot	5-shot	Avg
Qwen2 Instruct	7B	89.31	42.74	83.76	71.94
Gemma 1.1 it	7B	84.75	55.06	65.86	68.56
Aya-23	8B	73.62	33.80	77.05	61.49
Llama-3 Instruct	8B	85.77	78.57	68.22	77.52
Llama-3.1 Instruct	8B	88.97	89.25	87	88.41
PsychoLexLLaMA-pretrain-sft	8B	88.97	81.21	62.03	77.4
PsychoLexLLaMA-average	8B	90.10	89.03	90.04	89.72
Aya-23	35B	81.32	79.02	82	80.78
c4ai-command-r-v01	35B	87	78.06	75.08	80.05
Llama-3 Instruct	70B	90.55	88.58	76.77	85.3
Llama-3.1 Instruct	70B	93.02	92.63	92.1	92.58
PsychoLexLLaMA-pretrain-sft	70B	91.45	90.24	90.85	90.85
PsychoLexLLaMA-average	70B	92.13	91.85	91.87	91.95
Qwen2 Instruct	72B	91.11	73.79	92.29	85.73

Table 4 : Accuracy of LLMs on the PsychoLexEval dataset in English.

6.1.1 Performance Trends Across Models

The data reveal substantial variability in performance across models and settings. For instance, the Llama-3.1 Instruct with 70B parameters exhibits superior performance in all scenarios, suggesting a positive correlation between larger parameter sizes and enhanced comprehension and response accuracy. This trend is consistent in the English data, where models with larger parameters, such as Llama-3.1 Instruct 70B, also demonstrate robust performance, especially in zero-shot and five-shot settings.

Conversely, models with fewer parameters sometimes perform well in lower-shot settings but typically exhibit decreased performance as the complexity of tasks increases. For example, the Qwen2 Instruct with 7B parameters faces greater challenges in the Persian context than in English, potentially indicating linguistic or dataset-specific hurdles that are more effectively managed by larger models.

6.1.2 Language-Specific Observations

Our evaluation underscores distinct language-specific differences. In Persian, the increase in model accuracy from zero to five shots is more marked, indicating that Persian language models significantly benefit from added context. Conversely, English language models tend to have higher baseline performances, likely reflecting the advantages of more extensive pre-training datasets available in English.

6.1.3 Impact of Training and Fine-Tuning

The results particularly underscore the critical importance of targeted training and fine-tuning, as seen with the PsychoLexLLaMA models. Designed to surpass its predecessor, Llama 3.1, the 70B PsychoLexLLaMA occasionally does not reach its ambitious targets but consistently matches or exceeds the performance of the original Llama 3.1 model. This consistency indicates that while specific enhancements did not universally lead to improvements, they significantly bolstered the model's capabilities. The 70B version, with its vast parameter count, possesses the capacity to acquire a broader knowledge base, making it challenging to add new knowledge without forgetting previously learned information. Consequently, fine-tuning such a large model demands considerably more data to achieve better outcomes due to its complexity.

In contrast, the 8B version of PsychoLexLLaMA often outperforms larger models, suggesting that precise, domain-specific fine-tuning can yield remarkable effectiveness, even with fewer parameters. This success highlights the potential of smaller models, particularly when equipped with tailored enhancements for specific applications like psychological evaluations.

The varying impacts of scaling between the 8B and 70B versions suggest that while larger models possess a broad knowledge base enhancing their general performance, strategic fine-tuning is crucial for maximizing efficacy in specialized domains. This observation encourages further research into training strategies that optimize both large and small models for specific tasks, ensuring that they not only retain previous knowledge but also effectively integrate new information.

7 Conclusion

This study has significantly advanced our understanding of how large language models (LLMs) can be effectively tailored for applications within psychology. Through the integration of specialized psychological content, the development of the PsychoLexQA and PsychoLexEval datasets, and the creation of the PsychoLexLLaMA model, we have demonstrated the substantial benefits of targeted model training and fine-tuning.

Our findings indicate that specific pretraining and fine-tuning strategies substantially enhance the performance of LLMs in psychological settings, underscoring the critical role of thoughtful model architecture and training approaches. Notably, while larger models typically show strong performance, our results reveal that even smaller models can achieve exceptional outcomes when subjected to precise, domain-specific adjustments. This suggests a scalable potential for LLMs in psychological applications that can be adapted to different contexts and constraints.

In conclusion, this research not only sheds light on the current capabilities and challenges of using LLMs in psychology but also sets a foundation for future work. It encourages ongoing refinement of these models to improve their relevance and accuracy, thereby enhancing their utility in real-world psychological applications. Moving forward, we anticipate that continued advancements in model training methodologies and evaluation strategies will drive significant progress in the field, making

LLMs an indispensable tool in the arsenal of psychological research and practice.

Limitations

This study highlights the potential of customizing large language models (LLMs) for psychological applications, but several limitations may affect the generalizability of our findings. The PsychoLexQA and PsychoLexEval datasets, which are key to this research, contain inherent biases from the selected materials and question designs, restricting the application of results to broader psychological contexts. The reliance on freely licensed sources, particularly Hilgard's Introduction to Psychology (1953), limits the diversity and depth of psychological topics in this study. While the text provides a solid foundation in core principles, it does not reflect the significant advancements in psychology, such as developments in cognitive-behavioral therapy, neuropsychology, and cross-cultural psychology. Due to copyright restrictions, more recent, relevant texts could not be used, potentially limiting the model's engagement with modern psychological paradigms. Future iterations will aim to include up-to-date psychological literature to address this limitation and enhance the model's applicability to contemporary practice. Another major limitation is the need for high-performance hardware, such as GPUs, which can hinder replicability and limit access for researchers with fewer resources. Balancing model performance with minimal data and hardware remains challenging, often requiring trade-offs that reduce practical utility. Future research should aim to diversify training data and develop more resource-efficient methods to enhance the deployment of LLMs in psychology.

References

- Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, & Yadollah Yaghoobzadeh. (2024). Benchmarking Large Language Models for Persian: A Preliminary Study Focusing on ChatGPT. *ArXiv, abs/2404.02403*.
- Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, & Behrouz Minaei-Bidgoli. (2023). PersianLLaMA: Towards Building First Persian Large Language Model. *ArXiv, abs/2312.15713*.
- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, & Harkirat Behl. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Shrivats Agrawal. (2023). Are LLMs the Master of All Trades? : Exploring Domain-Agnostic Reasoning Skills of LLMs. *ArXiv, abs/2303.12810*.
- Anthropic. (2024,). The Claude 3 Model Family: Opus, Sonnet, Haiku.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, & Sebastian Ruder. (2024). Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.
- Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margaret Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron A. Hecht, Jeremy P. Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J Gross, & James W. Pennebaker. (2023). Using large language models in psychology. *Nature Reviews Psychology, 2*, 688 - 701.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, . . . Zhiwei Zhao. (2024). The Llama 3 Herd of Models.
- Omid Ghahroodi, Marzia Nouri, Mohammad V. Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, & Mohammad Hossein Rohban. (2024). Khayyam Challenge (PersianMMLU): Is Your LLM Truly Wise to The Persian Language? *ArXiv, abs/2404.06644*.
- Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiakuan Li, Bojian Xiong, & Deyi Xiong. (2023). Evaluating Large Language Models: A Comprehensive Survey. *ArXiv, abs/2310.19736*.

- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, & Weizhu Chen. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, *abs/2106.09685*.
- Luoma Ke, Song Tong, Peng Cheng, & Kaiping Peng. (2024). Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review. *ArXiv*, *abs/2401.01519*.
- Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, . . . Yadollah Yaghoobzadeh. (2020). ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, *9*, 1147-1162.
- Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, & Ziqi Wang. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *ArXiv*, *abs/2307.11991*.
- Shervin Minaee, Tomáš Mikolov, Narjes Nikzad, Meysam Asgari Chenaghlu, Richard Socher, Xavier Amatriain, & Jianfeng Gao. (2024). Large Language Models: A Survey. *ArXiv*, *abs/2402.06196*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, & Colin Raffel. (2023). Crosslingual Generalization through Multitask Finetuning. Annual Meeting of the Association for Computational Linguistics,
- OpenAI. (2023a). *Gpt-3.5*. <https://www.openai.com/>. Accessed: 2023-06-13.
- OpenAI. (2023b). *Gpt-4 technical report*. <https://arxiv.org/abs/2303.08774>
- Pedram Rostami, Ali Salemi, & Mohammad Javad Dousti. (2024). PersianMind: A Cross-Lingual Persian-English Large Language Model. *ArXiv*, *abs/2401.06466*.
- Oleh Shliakhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, & Tatiana Shavrina. (2022). mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, *12*, 58-79.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, & Juliette Love. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, & Sara Hooker. (2024). Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. *ArXiv*, *abs/2402.07827*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, & Hannaneh Hajishirzi. (2022). Self-Instruct: Aligning Language Models with Self-Generated Instructions. Annual Meeting of the Association for Computational Linguistics,
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, & Chi Wang. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, & Fei Huang. (2024). Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, & De-chuan Zhan. (2024). Continual Learning with Pre-Trained Models: A Survey. *ArXiv*, *abs/2401.16386*.

A Dataset samples

Figure 2 shows an example of PsychoLexQA dataset based on psychology document. Figure 3, 4, 5, 6 illustrates example of PsychoLexQA dataset based on self-instruct method. Figure 7 shows an example of the PsychoLexEval dataset that includes a multiple-choice question and answer. Each example is presented in English and Persian to enhance understanding.

Question: How do researchers measure brain activity during sleep and meditation?

Answer: Researchers measure brain activity during sleep and meditation using neuroimaging techniques such as electroencephalograms (EEGs), magnetic resonance imagery (MRI), and positron emission tomography (PET). These techniques record electrical changes in the brain, brain wave patterns, and activation of specific brain areas, allowing researchers to observe and analyze different states of consciousness.

سوال: محققان چگونه فعالیت مغز را در هنگام خواب و مدیتیشن اندازه گیری می کنند؟

پاسخ: محققان با استفاده از تکنیک‌های تصویربرداری عصبی مانند الکتروانسفالوگرام (EEG)، تصویربرداری تشدید مغناطیسی (MRI) و توموگرافی گسیل پوزیترون (PET) فعالیت مغز را در طول خواب و مدیتیشن اندازه گیری می‌کنند. این تکنیک‌ها تغییرات الکتریکی در مغز، الگوهای امواج مغزی و فعال‌سازی نواحی خاص مغز را ثبت می‌کنند و به محققان این امکان را می‌دهند تا حالت‌های مختلف هوشیاری را مشاهده و تجزیه و تحلیل کنند.

Figure 2: Example of PsychoLexQA dataset (Document-based).

Literature Review:

Instruct: Conduct a literature review on the effects of brain plasticity on learning and memory. Summarize key findings and theoretical themes.

Output: Literature review shows that brain flexibility significantly affects learning and memory. Most of the findings indicate that engaging in new activities causes neuronal growth, while stress can decrease flexibility. Theoretical themes emphasize the need for educational strategies that foster environments that foster learning and resilience.

دستور: یک مرور ادبیات درباره تأثیر انعطاف پذیری مغز بر یادگیری و حافظه انجام دهید. یافته های کلیدی و مضامین نظری را خلاصه کنید.

خروجی: مرور ادبیات نشان میدهد که انعطاف‌پذیری مغز به طور قابل توجهی بر یادگیری و حافظه تأثیرگذار است. عمده ی یافته ها حاکی از آن است که درگیر شدن در فعالیت های جدید باعث رشد نوروونی میشود، در حالی که استرس میتواند موجب کاهش انعطاف پذیری شود. مضامین نظری بر نیاز به استراتژی های آموزشی که محیط هایی را که یادگیری و تاب آوری را تقویت میکنند، مورد تأکید قرار میدهند.

Biological Psychology

Figure 3: Example of PsychoLexQA dataset (Self-instruct).

Experiment Design :

Instruct: Design an experiment to evaluate the effect of a growth mindset intervention on students' persistence in academic tasks. Include methods for measuring sustainability and outcomes.

Output: The experiment will include two groups of students: one group that will participate in a growth mindset intervention that includes workshops on the benefits of facing challenges and learning from failure, and a control group that will receive no specific intervention. Persistence is measured by the time spent on a challenging academic task and the number of tasks completed. The hypothesis is that compared to the control group, students in the growth mindset group show more persistence and performance in academic assignments and highlight the impact of mindset on learning behaviors.

دستور: یک آزمایش طراحی کنید تا تأثیر مداخله ذهنیت رشد را بر پایداری دانش‌آموزان در وظایف تحصیلی ارزیابی کنید. شامل روش‌هایی برای اندازه‌گیری پایداری و نتایج باشد.

خروجی: این آزمایش شامل دو گروه از دانش‌آموزان خواهد بود: یک گروه که در مداخله ذهنیت رشد شرکت می‌کنند که شامل کارگاه‌هایی درباره مزایای روبه‌رو شدن با چالش‌ها و یادگیری از شکست است و گروه کنترل که هیچ مداخله خاصی دریافت نمی‌کند. پایداری از طریق زمان صرف شده بر روی یک وظیفه تحصیلی چالش‌برانگیز و تعداد وظایف کامل شده اندازه‌گیری خواهد شد. فرضیه این است که دانش‌آموزان در گروه ذهنیت رشد در مقایسه با گروه کنترل پایداری و عملکرد بیشتری در وظایف تحصیلی نشان خواهند داد و تأثیر ذهنیت بر رفتارهای یادگیری را برجسته می‌کند.

Cognitive Psychology

Figure 4: Example of PsychoLexQA dataset (Self-instruct).

Scenario Response:

Instruct: In a scenario where a student has difficulty managing time, identify specific strategies to help him improve his skills.

Output: 1. Teach the student to create a prioritized to-do list to Organize tasks by urgency and importance 2. Introduce the concept of timing, where the student sets aside specific times for studying and taking breaks 3. Use digital tools such as calendars or time management apps to set reminders for deadlines. 4. Encourage regular reviews of their schedule so they can adjust and review their time management practices. 5. Promote self-evaluation of their time and productivity to identify and improve their patterns.

دستور: در سناریویی که دانش آموز در مدیریت زمان مشکل دارد، استراتژی های خاصی را برای کمک به او در بهبود مهارت های خود شناسایی کنید.

خروجی: 1. به دانش آموز بیاموزید که فهرست کارهای اولویت بندی شده ایجاد کند تا وظایف را بر اساس فوریت و اهمیت سازماندهی کند. 2. مفهوم زمان بندی را معرفی کنید، جایی که دانش آموز زمان های خاصی را برای مطالعه و استراحت اختصاص می دهد. 3. از ابزارهای دیجیتال مانند تقویم استفاده کنید. یا برنامه های مدیریت زمان برای تنظیم یادآوری برای مهلت ها. 4. بازبینی منظم برنامه خود را تشویق کنید تا بتوانند شیوه های مدیریت زمان خود را تنظیم و مرور کنند. 5. خود ارزیابی زمان و بهره وری خود را برای شناسایی و بهبود الگوهای آنها ارتقا دهید.

Cognitive Psychology

Figure 5: Example of PsychoLexQA dataset (Self-instruct).

Case Study Analysis:

Instruct: Analyze a 6-year-old boy who resorts to aggression in times of frustration. Explore possible causes and suggest strategies to manage his behavior.

Output: Case analysis of a 6-year-old boy who resorted to aggressive behaviors in times of frustration can reveal several potential factors. This behavior may be due to difficulty in regulating emotions, which may be related to the lack of coping strategies for frustration and anger. He may imitate aggressive behaviors he has observed in his environment, such as at home or in the media, or he may seek attention through these crises. Suggested strategies for managing his behavior could include teaching him emotion regulation techniques such as deep breathing and using 'I feel' sentences to express his feelings. Setting up a consistent behavior management plan, including clear rules and consequences, can help create structure. Finally, involving parents in reinforcing positive behavior at home and providing a loving environment can support the boy in constructively managing his frustrations.

دستور: یک پسر 6 ساله که در مواقع ناامیدی به تهاجم متوسل می‌شود را تحلیل کنید. علل احتمالی را بررسی کرده و استراتژی‌هایی برای مدیریت رفتار او پیشنهاد کنید.

خروجی: تحلیل مورد مربوط به پسر 6 ساله‌ای که در مواقع ناامیدی به رفتارهای تهاجمی متوسل می‌شود، می‌تواند چندین عامل بالقوه را نشان دهد. این رفتار ممکن است ناشی از دشواری در تنظیم احساسات باشد که ممکن است به عدم وجود استراتژی‌های مقابله‌ای برای ناامیدی و خشم مربوط باشد. او ممکن است شیوه‌های تهاجمی را که در محیط خود، مانند خانه یا رسانه‌ها مشاهده کرده است، تقلید کند یا ممکن است از طریق این بحران‌ها توجه جلب کند. استراتژی‌های پیشنهادی برای مدیریت رفتار او می‌تواند شامل آموزش تکنیک‌های تنظیم احساسات مانند تنفس عمیق و استفاده از جملات 'من احساس می‌کنم' باشد تا احساسات خود را ابراز کند. تنظیم یک برنامه مدیریت رفتار منسجم، از جمله قوانین و عواقب واضح، می‌تواند به ایجاد ساختار کمک کند. در نهایت، درگیر کردن والدین برای تقویت رفتار مثبت در خانه و فراهم کردن محیطی دوست‌داشتنی می‌تواند از پسر در مدیریت سازنده ناامیدی‌هایش حمایت کند.

Developmental Psychology

Figure 6: Example of PsychoLexQA dataset (Self-instruct).

Multiple-Choice QA:

Question: Which strategy is NOT considered a form of problem-focused coping?

Choices:

- 1) Defining the problem
- 2) Seeking emotional support
- 3) Generating alternative solutions
- 4) Changing personal goals

Correct Answer is 2.

سوال: کدام استراتژی به عنوان شکلی از مقابله متمرکز بر مشکل در نظر گرفته نمی‌شود؟

گزینه‌ها:

- 1) تعریف مسئله
- 2) جستجوی حمایت عاطفی
- 3) ایجاد راه حل‌های جایگزین
- 4) تغییر اهداف شخصی

پاسخ صحیح 2 است.

Figure 7: Example of PsychoLexEval dataset.

B Prompt template

In the below, we provide an example prompt for each task.

Generate a dataset with 40 detailed and comprehensive instructions for specific tasks in the subcategory defined below. The dataset should be structured with three components: input, and output. Each task should be available in both Persian and English. Inputs and outputs should be detailed and extensive.

Output Format:

```
[
{
  "id": 1,
  "Language": "English",
  "Subcategory": "Subcategory",
  "Task": "Task",
  "Input": "A clear and concise description of the task and specific details or data needed to perform the task if needed.",
  "Output": "The expected result or answer based on the input provided and should have long text."
},
{
  "id": 2,
  "Language": "Persian",
  "Subcategory": "زیرمجموعه",
  "Task": "تسک",
  "Instruction": "توضیح واضح و مختصر وظیفه",
  "Input": "اختیاری: جزئیات یا داده‌های خاص مورد نیاز برای انجام وظیفه",
  "Output": "نتیجه یا پاسخ مورد انتظار بر اساس ورودی ارائه شده"
},
...
]
```

Each language should have 20 rows (totaling 40 rows), covering the following tasks:

- Case Study Analysis
- Experiment Design
- Data Interpretation
- Scenario Response
- Theoretical Explanation
- Diagnostic Evaluation
- Ethical Dilemma Resolution
- Intervention Planning
- Literature Review
- Comparative Analysis

Ensure that the instructions, inputs, and outputs are diverse and cover a wide range of scenarios within each task type. and ensure to generate all 20 rows and not say it is a sample or something like this.

Subcategory:

```
[
"Clinical Psychology",
"Cognitive Psychology",
"Developmental Psychology",
"Social Psychology",
"Biological Psychology",
"Industrial-Organizational Psychology",
"Neuropsychology",
"Health Psychology",
"Educational Psychology",
]
```