A Psychological Strategy Annotation Method Using Multiple LLMs with a Chain of Thought Based on Deductive Reasoning

Jinran Wang*†, Jiaming Luo‡, Shaomeng Yang†§, Yongjie Zhou¶,
Xuefang Zhang*, Rongfeng Su†‖**, Nan Yan†‖** and Lan Wang†‖**
* Wuhan Research Institute of Posts and Telecommunications, China
† Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
‡ Shanghai Jiao Tong University, China
§ University of Chinese Academy of Sciences, Beijing, China
¶ Shenzhen Kangning Hospital, Shenzhen, Guangdong, China
‖ Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences, China
E-mail:{jr.wang2, sm.yang, rf.su, nan.yan, lan.wang}@siat.ac.cn, leojm2017@sjtu.edu.cn,
zhouyj2023@mail.sustech.edu.cn, zhangxuefang@fhxy.net.cn

***Abstract*—Using appropriate psychological counseling strategies is critical for improving the quality of large language model-driven mental health dialogues, yet a single large language model (LLM) often produces unreliable results due to ambiguous strategy definitions and degradation of thought in reasoning. Multi-LLM voting-based strategy annotation methods are often hampered in their overall decision-making performance due to the low accuracy of some individual models. To address this, we propose a multi-LLM system with a deductive-reasoning chain of thought. We built a dataset that comprises psychologist-client dialogue segments, each annotated with nine strategies by four professional psychologists. The system uses three LLMs as experts to generate initial strategy candidates,and then an LLM as a judge applies deductive reasoning—aligning dialogue evidence with strategy definitions—to annotate the optimal strategy. Experiments show that this approach outperforms single models, voting, and conventional chain-of-thought methods in F1 score and accuracy. This work advances LLM-based psychological strategy annotation, offering a reliable tool for mental health dialogue systems.**

I. Introduction

With the widespread use of large language models (LLMs) in mental health support dialogues, integrating professional psychological strategies is crucial to make model outputs resemble those of expert psychologists [1]. A common approach is fine-tuning LLMs on large datasets annotated with psychological strategies. However, annotating such datasets is time-consuming, labor-intensive, and requires psychological expertise, creating an urgent need for automated strategy annotation. LLMs excel in spoken language understanding and general comprehension, making them widely used for text summarization tasks [2], [3], [4]. Their natural language understanding and reasoning rival human abilities [5], making them suitable for annotation tasks often involving prompt engineering. For instance, prior studies have successfully annotated tweets for political stance or inferred authors' political leanings [6], [7]. Yet, single-model text annotation faces challenges: minor prompt changes can cause large output variations, and high randomness can lead to inconsistent results on repeated runs with the same input [8], [9]. In mental health support, strategy distinctions are often implicit in dialogue, requiring deep understanding and reasoning to extract, which amplifies single-model annotation challenges. Some studies explore combining multiple LLMs to improve task performance. Collaboration methods include debate, roundtable discussions, and voting mechanisms to reach a better consensus [10], [11]. Iterative architectures also exist, where each agent uses outputs from previous layers to generate responses [12]. For locally deployed LLMs, majority voting can outperform single models [13]. Multiple models usually require mechanisms to resolve conflicting results. Chain-of-Thought (CoT) integrates natural language reasoning into prompts [14], and CoT-based judges can optimize majority voting outcomes. However, research shows CoT suffers from "Degradation of Thought" (DoT), where LLMs struggle to self-correct initial errors [10]. When a single LLM annotates several strategies from nine strategies and uses CoT to eliminate wrong ones, DoT limits accuracy improvements. Inspired by prior work applying deductive reasoning by first identifying patterns inductively and then solving complex problems deductively [15], we designed a psychological strategy annotation system based on "deductive reasoning" CoT and multiple LLMs. Instead of a single model self-correcting, three "experts" LLMs pre-trained on different

**Corresponding authors.

This research is supported by National Natural Science Foundation of China (U23B2018), National Natural Science Foundation of China (NSFC 62271477), Shenzhen Science and Technology Program (JCYJ20220818101411025), Shenzhen Science and (JCYJ20220818102800001), Technology Shenzhen Program Science and Technology Program (JCYJ20220818101217037), Shenzhen Peacock Team Project (KQTD20200820113106007).

979-8-3315-7206-8/25/$31.00

datasets first narrow candidate strategies to 1-3. Then, the judge LLM applies deductive reasoning—a method for deriving necessary conclusions from premises —within its CoT to select the final strategy [16].

II. Psychological strategy annotation

Psychological strategy annotation refers to the systematic coding and analysis of counseling interactions to identify and categorize the strategies used by counselors and the reactions of clients. This approach helps in understanding the effectiveness of counseling techniques and how clients respond to them [17].

The main connection between mental health counseling and strategy annotation lies in the feedback loop derived from analyzing the counseling process. By annotating and examining the strategies employed by counselors alongside corresponding client responses, researchers can gain deeper insights into which techniques are most effective and under what circumstances [18]. This knowledge can then be used to improve LLM-based counseling practice and enhance overall counseling outcomes. Our study primarily focuses on annotating counselor strategies and provides a general methodological pathway that can be extended to the client behavior annotation.

III. Method

The overall architecture of the psychological counseling strategy annotation method is shown in Fig.1. Our system inputs include preprocessed dialogue texts and strategy definitions from the strategy knowledge base. The strategy annotation system consists of three LLMs acting as "experts" and one LLM as the "judge". Each expert model annotates a counseling strategy for each dialogue sample based on the psychologist's dialogue text and the strategy definitions, producing strategies A, B, and C. Then, the dialogue text along with the definitions of strategies A, B, and C are fed into the judge model, which produces the final strategy annotation result, saved as the Strategy Annotation Dataset.

Strategy annotation method design

A. Stage One: "Expert" LLM Strategy Annotation

In the first stage, three large language models act as "experts" and independently annotate a strategy for each dialogue sample using few-shot prompts. Assume the probabilities of the three models annotating the correct strategy are p_A, p_B and p_C.

B. Stage Two: "Judge" LLM Strategy Annotation

According to the Condorcet Jury Theorem, when $p > 0.5$, $P_{voting} > p(p = p_A, p_B, p_C)$ [19].

But what if one model has $p_A < 0.5$? In challenging tasks, some of the "experts" LLMs may underperform without our awareness. In that case P_{voting} may lower than p_B or p_C

The calculation shows that when $p_A = 0.4$, $p_B = 0.52$, $p_C = 0.53$, the theoretical value of the judge is approximately 0.141, which ensures that the probability of the final result being correct is not less than that of A, B, and C. We can mitigate poor model adverse effects and improve the overall system's strategy annotation accuracy by "intelligent select upon disagreement". So, after the first stage, we used a "judge" LLM. If all three "experts" LLMs agree, system adopt their answer. If the three models disagree, the "judge" LLM annotates one among them.

We propose constructing a deductive reasoning chain-of-thought for the "judge" LLM , guiding the model to focus on finding evidence that supports or refutes each candidate strategy.

Our prompt template provides the judge LLM with the following materials:

- Original Dialogue:

Counselor(Psychologist): So, what you understand is that when someone says this to you, you feel it is about your unique charm. In the past, you would think your younger brother was mischievous.

- Candidate Strategies Selected by "Expert" Models:

Emotional Reflection, Interpretation

- Strategy Definitions:

- Emotional Reflection: A type of listening technique where the counselor responds to the emotional parts of the client's information, including observing emotions, naming them, and repeating them back to the client. ...

- Interpretation: A type of influence technique in which the counselor, based on their intuition, identifies behaviors, patterns, goals, desires, and emotions that are implied or hidden in the client's expressed information. Interpretation responses target the implicit or even unconscious aspects of the client's information. ...

The judge LLM follows these deductive reasoning chain-of-thought steps:

1. Convert the provided strategy definitions in the prompt into major premises for deductive reasoning.
2. Extract minor premises from the dialogue text.
3. Draw a conclusion.

For each candidate strategy, the judge LLM evaluates it using either Affirmative Evidence Reasoning or Negative Evidence Reasoning, producing a positive or negative confidence score. After comparing the confidence scores of 2-3 strategies, the judge LLM annotates the strategy with the highest confidence as the final conclusion.

Below is a concrete example demonstrating how the LLM uses the deductive reasoning chain-of-thought:

Affirmative Evidence Reasoning

1. If the counselor repeats or paraphrases the client's statements in a different way (with the purpose of clarifying or confirming information), then this utterance uses "Emotional Reflection" (The model extracts the major premise from the strategy definition).

2. The counselor says "you feel", "you would think "-these phrases indicate the counselor is repeating the client's information of the emotional parts (The model finds key information in the dialogue text to establish the minor premise).

3. Therefore, this utterance uses "Emotional Reflection" (conclusion).

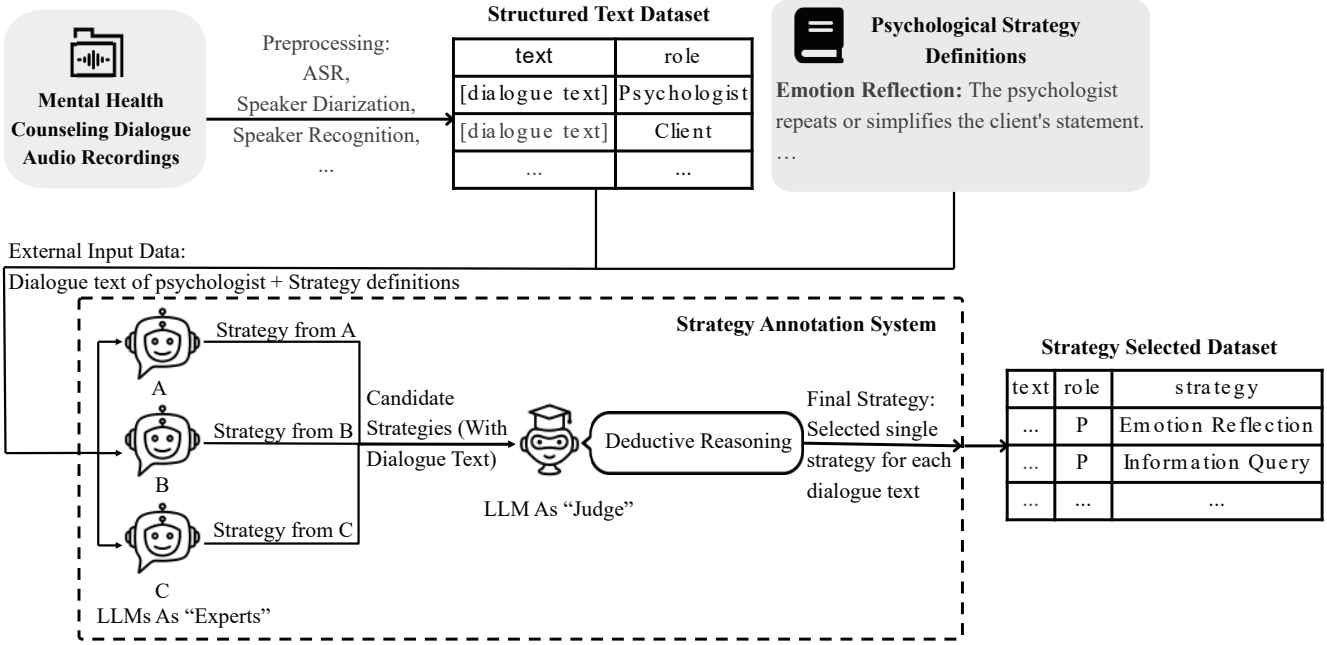

Fig. 1. The strategy annotation process with our strategy annotation system

Negative Evidence Reasoning

1. If the counselor provides a new perspective or interpretation (with the purpose of helping the client better understand their feelings, behaviors, or problems) (major premise, definition).

2. The counselor says "So, what you understand is"— these phrases indicate the counselor is confirming the client's viewpoint and not offering their own opinion, i.e., not providing a new perspective (minor premise, evidence).

3. Therefore, this utterance does not use "Interpretation" (conclusion).

IV. Experiments

A. Dataset

We collected approximately 300 hours of audio recordings of real Chinese psychologist–client counseling sessions. These recordings were transcribed into text using an automatic speech recognition (ASR) model based on FunASR [20], and speaker segmentation and identification were also performed. We further utilized a large language model to correct the transcribed text, fixing obvious typos and grammatical errors. Due to cost considerations, we sampled 266 psychologist dialogue text segments as the experimental data for this study.

B. Psychological Strategy Definitions

The strategy definition texts were transformed into standardized logical definitions to clearly convey the meaning of each strategy, with each defined using "truth definitions" [21].

Based on relevant studies, psychologists from our collaborating institution jointly discussed and defined nine strategies [22]. These strategies are listed in the following Table I, and their simplified forms were used in the Results section.

TABLE I
Strategies and their simplify

Strategy	Simplify
Challenge Confrontation	CC
Direct Guidance	DG
Emotion Reflection	ER
Emotional Support	ES
Explanation Provision	EP
Information Provision	IP
Information Query	IQ
Mindfulness Practice	MP
Self Disclose	SD

Four psychologists from a collaborating institution were invited to annotate the psychological strategy choices for this subset. We first calculated the consistency among the four human experts, using Cohen's kappa to compute the pairwise kappa and the average. The average kappa value among human experts on the sample set reached 0.635, exceeding the high consistency threshold of 0.6.

Among the strategy annotation results of four psychologists, the strategy chosen by the majority was considered correct. Specifically, if two or more psychologists annotate the same strategy, it is regarded as the correct choice. But samples where two psychologists choose one strategy and the other two choose another were deemed controversial and excluded. Only samples without disagreement were considered valid and were included in subsequent evaluations. In total, 244 of the

266 samples were valid.

C. System Construction

We deployed Qwen2.5-72B, Qwen2-72B, and Llama3-70B locally as expert models for Stage 1 annotation, and used DeepSeek-R1 as the judge model to perform deductive reasoning [23], [24], [25]. As an emerging open-source large language model, DeepSeek has been shown to have reasoning performance comparable to that of ChatGPT at a similar parameter scale [26], [27], [28].

D. Baseline Methods

Single Model

Using a single LLM to annotate the final strategy.The prompt template is as follows:

Input: Psychologist's dialogue text + definitions of nine strategies + instruction " Please annotate one strategy used by the psychologist in the dialogue based on the strategy definitions from the nine strategies."

Output: A single strategy label.

Majority Voting

We used majority voting with a deterministic tie-breaking rule to obtain the final strategy. When three single LLMs label strategies are A, B, and C respectively, if two or more agree on the same strategy, that majority result is taken as the final output. If all three differ, based on experience, the result from the strongest annotator model, Qwen2.5, is chosen as the final output.

Chain-of-Thought (CoT)

Using a common chain-of-thought instruction for the judge LLM. If all three LLMs agree on the same strategy, that result is taken as the final result. If there is disagreement, the results A, B, and C along with the strategy definitions are input into a judge LLM. The judge LLM's chain-of-thought was designed to: (i) briefly describe each strategy; (ii) summarize the dialogue; (iii) explain the reasoning; and (iv) provide the final strategy annotation.

E. Evaluation methods

We used F1 score and accuracy(ACC) as evaluation metrics. Due to class imbalance in the samples, we calculated a weighted F1 score based on the proportion of samples in each class.

V. Results

According to Table II results, the M-F, ISUD[1] method performs better than single-model baselines. Among the M-F, ISUD methods, the Deductive method (ours) achieves the best performance in both weighted F1 score and accuracy, reaching 0.664 and 0.668 ,respectively. Compared with the CoT method, the weighted F1 score of ours improves by approximately 5.9%, and accuracy improves by about 7.2%. Compared with the Voting method, the weighted F1 score of ours improves by about 21.6%, and the accuracy improves by about 22.6%.

Table III shows the F1 and accuracy for each category, and Table IV shows the sample counts of each category.

TABLE II
WEIGHTED AVERAGE F1 AND ACCURACY OF OURS AND BASELINE METHODS ON OUR 244 SAMPLES DATA SUBSET

Method		Weighted F1	Accuracy
M-F,ISUD	Deductive(Ours)	0.664	0.668
	CoT	0.627	0.623
	Voting	0.546	0.545
Single Model	Qwen2.5	0.542	0.537
	Qwen2	0.515	0.492
	Llama	0.499	0.500

Our method achieves significantly higher F1 and ACC scores than the models' native chain-of-thought and direct voting approaches in the CC, IP, and IQ categories. For DG, ER, ES, EP, and SD categories, our method performs slightly better or comparably to the baselines, yielding relatively stable results. However, in the MP category, the inclusion of the judge model reduces the final performance. Upon analyzing the original data, we found that the large models often misclassify many MP instances as other categories. Strategy statistics also show that two of the three expert models label fewer instances as MP. MP often involves the psychologist guiding the client through physical exercises like diaphragmatic breathing. Since our input does not provide the dialogue context for the psychologist's specific utterances, the lack of such background information makes it difficult for the model to determine whether the psychologist's "guidance" behavior belongs to the MP category. Additionally, we observed that the large models rarely label the "other" category, almost always annotating one of the nine defined strategies.

Overall, the results demonstrate that the "majority-first, intelligent select upon disagreement" strategy annotation methods significantly outperform single-model approaches. Among all intelligent annotation methods, ours deductive reasoning method achieves the best performance.

VI. Conclusions

Our study presents a psychological strategy annotation system that leverages multiple large language models combined with deductive chain-of-thought reasoning to enhance the accuracy of strategy annotation in psychological counseling dialogues. Initially, three "experts" models generate candidate strategies, followed by a "judge" model that applies deductive reasoning based on strategy definitions and dialogue context to annotate the most suitable counseling strategy. This design mitigates the instability of relying on a single model and addresses the degradation in chain-of-thought reasoning. When single-model accuracy is insufficient, the judge model improves the system's reliability.

Experimental results demonstrate that this multi-model collaborative system with deductive reasoning significantly improves both accuracy and F1 score on real psychological counseling dialogue datasets.

[1]We call the method that uses multiple LLMs to generate initial strategy candidates and then decides the final strategy through a majority vote with intelligent selection upon disagreement "majority-first, intelligent select upon disagreement."(M-F, ISUD).

TABLE III
F1 AND ACCURACY OF OURS AND BASELINE METHODS ON EACH CATEGORY

Method			Deductive	CoT	Voting	Qwen2.5	Qwen2	Llama3
Strategy	CC	F1	**0.700**	0.537	0.444	0.439	0.450	0.526
		ACC	**0.875**	0.688	0.625	0.562	0.562	0.625
	DG	F1	**0.711**	0.692	0.492	0.559	0.533	0.557
		ACC	**0.771**	**0.771**	0.457	0.486	0.457	0.457
	ER	F1	0.400	**0.500**	0.316	0.326	0.167	0.378
		ACC	0.462	0.461	0.462	**0.538**	0.462	0.462
	ES	F1	**0.654**	0.632	0.571	0.512	0.476	0.431
		ACC	0.739	**0.783**	0.609	0.478	0.652	0.609
	EP	F1	**0.542**	0.522	0.522	0.524	0.458	0.435
		ACC	0.481	0.444	0.444	0.407	0.407	**0.556**
	IP	F1	**0.724**	0.615	0.517	0.549	0.538	0.456
		ACC	**0.636**	0.485	0.455	0.515	0.424	0.394
	IQ	F1	**0.800**	0.689	0.667	0.576	0.630	0.603
		ACC	**0.774**	0.677	0.613	0.548	0.548	0.613
	MP	F1	0.761	0.714	**0.808**	**0.808**	0.773	0.606
		ACC	0.696	0.652	**0.913**	**0.913**	0.739	0.435
	SD	F1	0.821	**0.840**	0.692	0.694	0.638	0.723
		ACC	**0.852**	0.778	0.556	0.630	0.556	0.630
	OTHER	F1	0.154	**0.270**	0.154	0.138	0.125	0.000
		ACC	0.125	**0.313**	0.125	0.125	0.125	0.000

TABLE IV
SAMPLE COUNTS OF EACH CATEGORY

Method		Ground truth	Deductive	CoT	Voting	Qwen2.5	Qwen2	Llama3
Strategy count	CC	16	24	25	29	25	24	22
	DG	35	41	43	30	33	25	26
	ER	13	17	11	25	30	35	24
	ES	23	29	34	26	20	40	42
	EP	27	21	19	19	15	21	42
	IP	33	25	19	25	29	19	24
	IQ	31	29	30	26	28	23	32
	MP	23	19	19	29	29	21	10
	SD	27	29	23	25	22	20	20
	OTHER	16	10	21	10	13	16	2

Additionally, we observed that when nine strategy categories are defined, the models tend to annotate from these categories rather than choosing "other." Future work will investigate this tendency and improve the system for cases without a correct answer. Specifically, if none of the three expert models annotate the correct strategy, the judge model can autonomously choose the correct one to enhance accuracy.

REFERENCES

[1] H. Qiu, A. Li, L. Ma, and Z. Lan, "Psychat: A client-centric dialogue system for mental health support," in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2024, pp. 2979–2984.

[2] M. He and P. N. Garner, "Can chatgpt detect intent? evaluating large language models for spoken language understanding," *arXiv preprint arXiv:2305.13512*, 2023.

[3] G. Li, L. Chen, and K. Yu, "How chatgpt is robust for spoken language understanding?" In *Proceedings of Interspeech*, 2023, pp. 2163–7.

[4] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng, "Exploring the limits of chatgpt for query or aspect-based text summarization," *arXiv preprint arXiv:2302.08081*, 2023.

[5] J. Yang et al., "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2024.

[6] K. Joseph, L. Friedland, W. Hobbs, O. Tsur, and D. Lazer, "Constance: Modeling annotation contexts to improve stance classification," *arXiv preprint arXiv:1708.06309*, 2017.

[7] P. Törnberg, "Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning," *arXiv preprint arXiv:2304.06588*, 2023.

[8] M. V. Reiss, "Testing the reliability of chatgpt for text annotation and classification: A cautionary remark," *arXiv preprint arXiv:2304.11085*, 2023.

[9] R. D. Kristensen-McLachlan, M. Canavan, M. Kárdos, M. Jacobsen, and L. Aarøe, "Are chatbots reliable text annotators? sometimes," *PNAS nexus*, vol. 4, no. 4, pgaf069, 2025.

[10] T. Liang et al., "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint arXiv:2305.19118*, 2023.

[11] J. C.-Y. Chen, S. Saha, and M. Bansal, "Reconcile: Round-table conference improves reasoning

via consensus among diverse llms," *arXiv preprint arXiv:2309.13007*, 2023.
[12] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou, "Mixture-of-agents enhances large language model capabilities, 2024," *URL https://arxiv.org/abs/2406.04692*,
[13] J. Niimi, "Dynamic sentiment analysis with local large language models using majority voting: A study on factors affecting restaurant evaluation," *arXiv preprint arXiv:2407.13069*, 2024.
[14] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
[15] C. Cai et al., "The role of deductive and inductive reasoning in large language models," *arXiv preprint arXiv:2410.02892*, 2024.
[16] R. DeWitt, *Worldviews: An introduction to the history and philosophy of science*. John Wiley & Sons, 2018.
[17] A Li et al., "Understanding client reactions in online mental health counseling. arxiv," *arXiv preprint arXiv:2306.15334*, 2023.
[18] V. Pérez-Rosas, X. Sun, C. Li, Y. Wang, K. Resnicow, and R. Mihalcea, "Analyzing the quality of counseling conversations: The tell-tale signs of high-quality counseling," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
[19] D. M. Estlund, "Opinion leaders, independence, and condorcet's jury theorem," *Theory and Decision*, vol. 36, no. 2, pp. 131–162, 1994.
[20] Z. Gao et al., "Funasr: A fundamental end-to-end speech recognition toolkit. arxiv 2023," *arXiv preprint arXiv:2305.11013*,
[21] Y. Zhao, *Logic Tutorial*. China Renmin University Press, 2014.
[22] C. E. Hill, "Helping skills: Facilitating exploration, insight, and action," *American Psychological Association*, 1999.
[23] A. Y. Qwen et al., "Qwen2. 5 technical report," *arXiv preprint*, 2024.
[24] J. Bai et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
[25] A. Dubey et al., "The llama 3 herd of models," *arXiv e-prints*, arXiv–2407, 2024.
[26] R. Phogat, D. Arora, P. S. Mehra, J. Sharma, and D. Chawla, "A comparative study of large language models: Chatgpt, deepseek, claude and qwen," in *2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, IEEE, 2025, pp. 609–613.
[27] I. Jin et al., "Deepseek vs. chatgpt: Prospects and challenges," *Frontiers in Artificial Intelligence*, vol. 8, p. 1 576 992, 2025.
[28] M. N.-U.-R. Chowdhury, A. Haque, and I. Ahmed, "Deepseek vs. chatgpt: A comparative analysis of performance, efficiency, and ethical ai considerations," *Authorea Preprints*, 2025.